NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

CATEGORIZATION OF SURVEY TEXT UTILIZING NATURAL LANGUAGE PROCESSING AND DEMOGRAPHIC FILTERING

by

Christine M. Cairoli

September 2017

Thesis Advisor: Lyn R. Whitaker
Second Reader: Andrew T. Anglemyer

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704–0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2017	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE CATEGORIZATION OF SURVEY TEXT UTILIZING NATURAL LANGUAGE PROCESSING AND DEMOGRAPHIC FILTERING			**5. FUNDING NUMBERS**
6. AUTHOR(S) Christine M. Cairoli			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number NPS.2017.0015-IR-EP5-A.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

Thousands of Navy survey free text comments are overlooked every year because reading and interpreting comments is expensive, time consuming, and subjective. Valuable information from these comments is not being utilized to make important Navy decisions. We provide a new procedure to automate the identification of primary topics in short, jargon laced, topic based survey comments by applying a label to each comment and then using those labels to bin comments into operationally meaningful categories. We apply this method to the Navy Retention Survey to provide the Chief of Naval Personnel with an objective analysis of the questions "Why are sailors leaving?" and "What will make sailors stay on active duty?" Furthermore, we introduce an implementation of this method using the Demographic Analysis of Responses Tool for Surveys (DARTS), which allows us to filter comment bins using the over 100 demographic and military status elements associated with each sailor. By targeting critically undermanned specialties, the reports generated with this tool provide quantifiable results that allow retention policy makers the ability to review, modify, and create relevant incentives to retain critically talented sailors to meet fiscal year end strength and operational requirements.

14. SUBJECT TERMS Navy retention, survey comments, comment labeling, text analysis, natural language processing			**15. NUMBER OF PAGES** 91
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

CATEGORIZATION OF SURVEY TEXT UTILIZING NATURAL LANGUAGE PROCESSING AND DEMOGRAPHIC FILTERING

Christine M. Cairoli
Lieutenant, United States Navy
B.S., United States Naval Academy, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2017

Approved by: Lyn R. Whitaker, Ph.D.
Thesis Advisor

Andrew T. Anglemyer, Ph.D.
Second Reader

Patricia A. Jacobs, Ph.D.
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Thousands of Navy survey free text comments are overlooked every year because reading and interpreting comments is expensive, time consuming, and subjective. Valuable information from these comments is not being utilized to make important Navy decisions. We provide a new procedure to automate the identification of primary topics in short, jargon laced, topic based survey comments by applying a label to each comment and then using those labels to bin comments into operationally meaningful categories. We apply this method to the Navy Retention Survey to provide the Chief of Naval Personnel with an objective analysis of the questions "Why are sailors leaving?" and "What will make sailors stay on active duty?" Furthermore, we introduce an implementation of this method using the Demographic Analysis of Responses Tool for Surveys (DARTS), which allows us to filter comment bins using the over 100 demographic and military status elements associated with each sailor. By targeting critically undermanned specialties, the reports generated with this tool provide quantifiable results that allow retention policy makers the ability to review, modify, and create relevant incentives to retain critically talented sailors to meet fiscal year end strength and operational requirements.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ACM	Association for Computing Machinery
AC	active component
AFO	absolute first occurrence
BO	BusinessObjects
BUPERS-3	Military Community Management (BUPERS-3)
CIMS	Career Information Management System
CIP	Career Intermission Program
CNO	Chief of Naval Operations
CNP	Chief of Naval Personnel
CTS	candidate token score
CVSS	Career Viewpoint Surveys and Studies
C-WAY	Career Waypoints
DARTS	Demographic Analysis of Responses Tool for Surveys
DCNO	Deputy Chief of Naval Operations
DOD	Department of Defense
DTM	document term matrix
ESR	electronic service record
EDLN	estimated date of loss to the navy
Freq	frequency
FH	first half
FQ	first quarter
FTS	full time support
GUI	graphical user interface
LDA	Latent Dirichlet Allocation
MILPERSMAN	*Naval Military Personnel Manual*
MSR	minimum service requirement
N1	Manpower, Personnel, Education and Training
N13	Military Personnel Plans and Policy Division
NAVADMIN	Navy-specific administrative message
NPRST	Navy Personnel Research, Studies, and Technology

NSIPS	Navy Standard Integrated Personnel System
OPNAV	Office of the Chief of Naval Operations
PDF	portable document format
RC	reference commonness
RFO	relative first occurrence
POS	parts of speech
PRD	projected rotation date
PTT	partial technical terms
SEAOS	soft end of active obligated service
SPAWAR	Space and Naval Warfare
SRB	selected reenlistment bonus
TS	token size
TT	technical term
UIC	unit identification code
URL	unrestricted line
VBA	Visual Basic for Applications

EXECUTIVE SUMMARY

From July 2016 to July 2017, individual Navy organizations administered 96 officially authorized Navy surveys through the self-service web-based survey application, MAX.gov. Approximately 70% of the authorized surveys contain at least one text comment box that must be individually read for the information to be available (R. Linton, personal communication, August 28, 2017). Reading and interpreting comments is expensive, time consuming, and subjective, and therefore, thousands of Navy survey comments are overlooked. Valuable information from these comments is not being utilized to make important Navy decisions.

To assist the fleet in analyzing free-text comment response to survey questions, our research provides a new procedure to automate the identification of primary topics in short, jargon-laced, topic-based survey comments by applying a label to each comment. The labels are then used to bin comments into operationally meaningful categories. The comment analysis method we develop is a general method that can be applied to any set of short comments and has two steps.

The first step, based on work by Chuang et al. (2012) assigns a 1- to 3-word label to each comment. This process starts by preprocessing comments to ensure proper formatting. Then we assign each comment a set of candidate token that includes all 1- to 3-word consecutive word combinations from the comment. The candidate tokens are assigned statistical and linguistic variables of two types, token-specific variables and comment-specific variables. These variables are used to construct a candidate token score (CTS) for each candidate token.

Token-specific variables are unique to each candidate token and independent of the comments associated with the token. The four token specific variables are calculated once for each unique candidate token in the corpus and are then factored into the final CTS computation. The token size (TS) is the categorical factor of the number of words in the token. Chuang et al. (2012) defines technical terms as a multi-word phrase that meets a specific pattern; it begins with either an adjective or noun, strings together adjectives,

nouns, or prepositions in the middle, and ends in a noun. The variable TT indicates whether the token is a technical term or not. Chuang et al. (2012) further defines partial technical terms as tokens that match a substring of a technical term. Our variable, PTT, is an indicator of whether a token is a partial technical term or not. Reference commonness (RC) uses a reference corpus with terminology, an acronym, and language use that is consistent with the corpus of short-text comments. We stem the reference corpus to account for occurrences of different-tense words of the same concept and then we calculate the RC for each token contained in the reference corpus by dividing the log of the token frequency in the reference corpus by the log of token frequency of the most frequent token of the same token size. The RC is then assigned to a six-level categorical variable that is divided based on the following sets: {0}, (0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], and (0.8, 1.0].

Comment-specific variables are unique to each comment and must be calculated for each candidate token-comment pair. For surveys with many comments, the processes to compute these variables need to be succinct and efficient. The three comment-specific variables are Freq, the frequency of the token in each comment; RFO, a measure of the first occurrence of token relative to a token of the same frequency; and FH, an indication of whether a token is contained in the first half of a comment or not.

Once all the necessary variables are constructed, the CTS is calculated for each candidate token and comment combination as a linear function of those variables using regression coefficients estimated using the approach of Chuang et al. (2012). Using the essence of this work, we randomly select fifty comments that include at least five words from each of four questions across three surveys to construct our dataset using two questions each from the Navy Retention Survey (Navy Standard Integrated Personnel System, CVSS, n.d.) and 2017 Female Dress Uniform & Cover Survey conducted by OPNAV N1 (MAX.gov, 2017). These four questions cover the two topics of manpower and uniforms and the Department of the Navy's *Naval Military Personnel Manual* (MILPERSMAN) (2002) and Navy Uniform Regulations (Department of the Navy, 2011) are used for the reference copra. We read the comments to determine the primary 1- to 3-gram label for each comment and store them as the expert label. We create all

1- to 3-gram tokens for each comment, excluding the expert label, and randomly select ten tokens to use as negative responses with a weight of 0.1. This produces a data set of 2,200 comments with labels. We determine the variable values for each label comment pair and fit a logistic regression to estimate our own regression coefficients. Our coefficients are shown in Table 1.

Table 1. Regression Coefficients for Candidate Token Score Calculation

Model Variable	Coefficient Estimate	Standard Error
(Intercept)	-2.5217800 ***	0.6249
TS ϵ {2}	-1.1246629 **	0.4086
TS ϵ {3}	-1.2805272 **	0.4512
TT	3.2928379 ***	0.5675
PTT	-1.0478745 *	0.4384
RC ϵ (0%, 20%]	-0.2783750	0.5632
RC ϵ (20%,40%]	-0.8293879 •	0.4572
RC ϵ (40%, 60%]	-0.5307969	0.4456
RC ϵ (60%, 80%]	-0.1845462	0.5033
RC ϵ (80%, 100%]	-2.9243750 *	1.1700
Log(Freq)	0.5744200	0.8267
RFO	3.8014855 ***	0.8273
FH	0.3303841	0.5082

Statistical significance = ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, •: $p < 0.1$

The candidate tokens for each comment are scored using the coefficients from Table 1. The candidate token with the maximum CTS among candidate tokens for a comment is assigned to be the label for that comment.

These labels on their own provide a summary of the primary topics in the comments. However, there are typically too many of them to provide a meaningful summary of the important primary topics. Taking our analysis one step further, the

second step of our process uses the labels and a systematic approach to group comments with similar primary topics into bins using more traditional visual text mining methods including Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). We use the resulting themes to choose prevalent words or phrases, known as keywords, to assign to each observable theme from the comments. We compare keywords to the comment labels and original text to group comments into topic bins. In this step, the analyst can use their subject matter expertise to modify binning and descriptions of those bins. Assigned comment bins are then displayed in tabular and graphical formats, which may be filtered to display results for specific populations.

Furthermore, we introduce an implementation of this method, our Demographic Analysis of Responses Tool for Surveys (DARTS). This tool allows us to filter comment bins using the over 100 demographic and military status elements associated with each sailor. By targeting critically undermanned specialties, the reports generated with this tool provide quantifiable results that allow retention policy makers the ability to review, modify, and create relevant incentives to retain critically talented sailors to meet fiscal year end strength and operational requirements.

We apply our comment analysis methodology to the Navy Retention Survey to provide the Chief of Naval Personnel (CNP) with an objective analysis of the questions "Why are sailors leaving?" and "What will make sailors stay on active duty?" Our results find that naval officers are leaving primarily because of civilian career opportunities and that increased promotion will help with retention. Filtering on Unrestricted Line (URL) female officers in their first tour where retention has been a congressional focus has proven that over 30% of this group leaves active duty because they feel they spend too much time away from home, while 25.6% indicate that they wish to start or focus more on their family. Members of this same demographic comment that better work-life balance and more family time would encourage them to remain on active duty. Although bonuses are often pushed at this population, only 3% indicate that monetary compensation will encourage them stay. With these quantifiable results, retention policy makers are better able to review, modify, and create more relevant incentives to retain

"our best sailors" while working within budget constraints and meeting fiscal year end strength and operational requirements.

Our method is generalizable beyond the realm of Navy retention. With over 65 surveys a year containing short comments, survey administrators are always seeking tools to expedite the evaluation process and utilization of the comments. This research opens the door for a more effective feedback loop by analyzing more comments in a shorter period of time. With information from the comments, leadership can respond to sailor concerns and demonstrate the value of completing surveys. If sailors see that surveys make a direct impact, they will be more willing to complete them, continuing the chain of improving effective communication throughout the Navy.

Furthermore, our methods can be adapted to non-Navy surveys. The ability to adapt our approach by using a context-specific reference corpus from any manual, document, or website allows the methodology to be applied to other surveys with a different set of jargon and acronyms. Because of this, our method is easily adaptable to any survey with topic based comments.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Chuang, J., Manning, C., & Heer, J. (2012). Without the clutter of unimportant words. *ACM Transactions on Computer-Human Interaction (TOCHI), 19*(3), 1–29. doi:10.1145/2362364.2362367

Department of the Navy. (2002). *Naval military personnel manual* (MILPERSMAN) (NAVPERS 15560D). Millington, TN: Bureau of Naval Personnel.

Department of the Navy. (2011). United States Navy uniform regulations (NAVPERS 15665I). Millington, TN: Bureau of Naval Personnel.

MAX.gov. Surveys. (2017). *2017 Female dress uniform & cover survey*. Retrieved from https://survey.max.gov

Navy Standard Integrated Personnel System, CIMS. (n.d.). *Career viewpoint retention survey*. Retrieved from https://nsipsprod.nmci.navy.mil/

Navy Standard Integrated Personnel System, CVSS. (n.d.). *Career viewpoint retention survey*. Retrieved from https://nsipsprod.nmci.navy.mil/

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Lyn Whitaker, for graciously agreeing to tackle a topic that was not standard in our curriculum. From the beginning, you were supportive and provided the guidance needed to learn new areas of data analysis that made this thesis possible. I appreciate your addition of a directed study class during your off quarter and accommodating my unscheduled visits, where you stopped your work to answer my questions. You let me continue to learn through all of my "scope creep," which provided better results and methodology. I had great days and others that kept me away from school for extended periods of time. Throughout it all, you provided me the encouragement that I needed to continue; I would not have finished without you!

I would also like to thank my second reader, Dr. Andy Anglemyer. I appreciate that you took on my thesis late in the process and accommodated my numerous draft revisions. You provided the necessary survey insight that would not have been addressed otherwise.

This research was also made possible with the support of OPNAV N1, particularly Dr. Richard Linton. I extend sincere gratitude for working with me to obtain the approval to use all of the data, answer my numerous questions, and provide continued liaison support. I appreciate the support of Wayne Wagner for the time and effort to review and validate data. Thank you to the OPNAV N1 codes that provided feedback and use of their survey responses, allowing me to improve and validate my methodology.

My journey at NPS had many challenges, but the continued support of many people ensured my success. I would like to thank those professors who increased my eagerness to learn and provided additional support along the way: Professors Buttrey, Carlyle, Huddleston, Hyink, McLemore, and Tick. Finally, thank you to classmates Jason Pfaff, Bill Buffington, Joe Fleshman and Bill Huff. You have all provided support that improved this research, directly or indirectly, and I am grateful that you were here at NPS to help me in my journey.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

From July 2016 to July 2017, individual Navy organizations administered 96 officially authorized surveys through the self-service web-based survey application, found at MAX.gov. Approximately 70% of the authorized surveys contain at least one text comment box. These include responses to some of the Navy's most pressing questions such as, "What would make you stay on active duty?" Many of these comment boxes are overlooked since reading and interpreting comments is expensive, time consuming, and subjective (R. Linton, personal communication, August 28, 2017). With 100 to 2500 responses to each survey, there are indications that thousands of unread comments are available that might have a bearing on congressional inquiries, top leadership initiatives, command-level issues, and naval-manning shortfalls. With the conversion of the survey process to a self-service system, any Navy service member can be tasked with conducting and analyzing these surveys and, specifically, the comment boxes with no tools available for the text analysis. The purpose of this research is to assist the fleet in analyzing free-text comment response to survey questions. We automate the identification of primary topics in survey comments by applying first a label to each comment. We then use these labels to group comments into a few meaningful categories to provide objective, quantifiable results for survey questions with text responses.

A. BACKGROUND MOTIVATION

The Chief of Naval Personnel (CNP) is the Navy's three-star admiral responsible to the Chief of Naval Operations (CNO) for overall manpower readiness of the Navy. One of his top priorities is talent management and associated initiatives to retain the Navy's top performing sailors while ensuring proper manning across all job specialties. These efforts, in addition to increased pressure from Department of Defense (DOD) leadership to minimize costs, are pertinent to ensure a diverse workforce while meeting all Navy requirements. An emphasis is placed on using the Navy Retention Survey to determine how best to shape the Navy and retain top performing sailors while being able to justify the need for certain policies, such as the career intermission program (CIP) or

selective reenlistment bonuses (SRB). The Navy Retention Survey offers direct fleet feedback and to align retention efforts with sailor needs and desires (R. Linton, personal communication, October 14, 2016).

Navy surveys contain questions of various types including those with comment boxes for a free-text response. These are the least reviewed components of a survey and yet these free-text responses provide the most direct insight into sailor retention. A primary example is the responses to why sailors are leaving the Navy, where the most frequent reason is reported as *Other*, with over 20% of the responses. This question has a comment box that asks for *Other* responses to be clarified, but no analysis has been conducted to explain these reasons. Another important question asks sailors, "What can be done to encourage you to remain in the Navy on active duty when you are next required to make a stay/leave decision?" This comment box has generated 13,781 responses in three years and has not been reviewed or considered when making policy recommendations to increase Navy retention. Conducting an in-depth analysis of these comment boxes can provide invaluable information to CNP and his analytic team to help retain critical personnel.

B. OVERVIEW

We develop an algorithm that automatically labels and then, with some analyst input, classifies the text of Navy Retention Surveys into generalized topics. This method allows for more timely analyses and implementation of actions based on those analyses. The most popular approaches for identifying common topics among text documents of a corpus are in the class of clustering algorithms called topic models, which include Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) and the more recent work of label selection (Chuang, Manning, & Heer, 2012b). In our setting, text comments are documents, and the corpus is the set of all comment responses for a question of a specific survey. Most of the successful applications of topic modeling methods are for longer documents, including dissertations, journal articles (Chuang et al., 2012b), and news articles (Mei, Shen, & Zhai, 2007). Navy survey comment boxes do not fit into this category and range in size from a single word (e.g., "nothing") to small paragraphs with

approximately 1000 characters. Nitin, Swapna, & Shankararaman (2015) suggest methods of using agglomerative clustering and sentiment analysis techniques to analyze shorter survey comments. These methods require extensive analyst setup for each comment, and with so many surveys and limited time, they are impractical for Navy surveys.

Furthermore, Navy comments often include Navy-specific acronyms and jargon that are difficult to associate with the same topic even though they have the same meaning, especially when the terms are misunderstood by sailors. An example of this is the set of terms "HYT," "higher tenure," "high year tenure," and "higher tenor," which corresponds to the official term "high year tenure," a force management policy restricting the length of service based on a member's paygrade (Department of the Navy, 2002). These terms are related to the term "failed officer selection," where an officer fails to promote to the next higher paygrade in the required number of attempts. Together, these terms, along with many other related terms, fall under the topic "continued service not authorized." With so many terms that are associated with each other only because of their naval context, typical text analytic methods do not provide timely, efficient, and reliable results.

To overcome the limitations of current text analytic methods, we propose a methodology that adapts previous research to work on contextual based, short-text comments. Our method uses cautious text preprocessing to reduce easily identifiable duplicates with acronym and contraction substitution. We then use a two-step approach for identifying and classifying comments into a few meaningful primary topic categories rather than use topic models or other clustering methods that are more common. In the first step of our two-step approach, we assign a label to each comment using an approach inspired by Chuang et al. (2012b). We modify their method by adapting it to our smaller amount of text. Additionally, we rely on readily available reference documents or webpages that provide the contextual link and help identify terms that are most related to the comment. With a descriptive label assigned to each comment, we take our method a step further and outline a systematic approach for using these labels to bin comments

with similar topics. We use a combination of the usual text analytic method on the labels with some, but minimal, analyst input for topic discovery and validation.

To make the binned primary topics more useful and easily reviewed, we develop and illustrate a user-friendly Demographic Analysis of Responses Tool for Survey (DARTS) to provide results filtered by combinations of over 100 military-related and demographic fields such as paygrade, years of service, gender, and military community. This general tool is adaptable to any survey with demographic fields, but is developed for analysis of the Navy Retention Surveys to provide the Office of the Chief of Naval Operations (OPNAV), Deputy Chief of Naval Operations for Manpower, Personnel, Education and Training (N1) with an easy method to filter their survey responses by critical demographics.

C. BENEFITS OF THE STUDY FOR THE NAVY

In 2016, government downsizing and budget cuts caused the disestablishment of Navy Personnel Research, Studies, and Technology (NPRST), which included survey experts and analysts who conducted the Navy's mandated surveys. The United States Navy is still required to conduct congressionally mandated and DOD directed surveys, but these responsibilities are farmed out to individual organizations with little or no survey expertise or analytic background. Many of the surveys and their analyses are conducted using tools available at the MAX.gov website, but this website has no available resources for text analysis. Our method provides a resource that can be applied to any short topic-based comment. It quickly identifies prevalent signal topics to determine whether further comment review is needed. It is used to provide objective responses to "short fused" tasking with available survey data and is easily adaptable with a valid reference document. These aspects result in significant time and cost savings for analysts and the ability to routinely incorporate survey comment responses. As more comments are reviewed, the results can be conveyed back to the fleet, showing leadership response and proving to sailors that their feedback is being heard. Seeing such results encourages sailors to provide more feedback, which results in more survey participation. This feedback loop is critical to maintaining strong communication throughout the fleet.

Additionally, as fiscal constraints continue to impact policy makers, Navy leadership must find alternative methods to ensure adequate sailor retention while maintaining diversity and equal opportunities. The application of our method to the Navy Retention Survey provides OPNAV N1, and his staff with fleet feedback that categorizes the most effective ways to retain high-quality sailors. With DARTS, all survey results, including comments, can be filtered and grouped by demographic, naval community, length of service, and over 100 other factors to target specific retention areas that typically have low retention rates. As a result of our work, decision makers can now have timely and critical information that can directly affect annual end strength and fleet operational manning.

D. ORGANIZATION OF THESIS

The remaining chapters provide a detailed account of our approach and how survey comments are analyzed using it. Chapter II discusses the methodology of our model that is used to create comment labels and assign them to topic bins. Chapter III applies our method the Navy Retention Survey to answer the questions, "Why are sailors leaving active duty?" and "What will encourage sailors to remain on active duty?" In Chapter IV, we validate our methodology by comparing our results to those of Navy analyst. Chapter V provides a conclusion, and recommendations for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

II. METHODOLOGY

The comment analysis method we develop is a general method that can be applied to any set of short comments and has two steps. The first uses a modification of work by Chuang et al. (2012b) to assign a 1- to 3-word label to each comment. These labels on their own provide a summary of the primary topics in the comments. However, there are typically too many of them to provide a meaningful summary of the important primary topics. Taking the analysis one step further, the second step uses the labels and a systematic approach to group comments with similar primary topics into bins. In this step, the analyst can use their subject matter expertise to modify binning and descriptions of those bins. Assigned comment bins are then displayed in tabular and graphical formats which may be filtered to display results for specific populations.

A. COMMENT LABEL

1. Preprocess Data

To make the method as robust as possible, we take a minimalist approach to preprocessing the text. The preprocessing steps include removing punctuation except for those that are needed for parts of speech tagging, including periods, commas, semi-colons, and colons. We convert the text to lower case and convert common contractions to their whole word equivalent. We do not remove stop words such as "and," "the," "a" since the comments are short and the stop words may add necessary descriptors. We also do not stem words. Stemming converts words to their root form, removing any inflection so that the fundamental meaning of the word is captured (Zhai & Massung, 2016). For comments that include many acronyms or words with similar meanings, we construct a lexicon with a substitution word for each theme. Different lexicons will be required for different surveys. A Navy example of terms that need substitutions include the phrases "failed officer selection," "failure to select," "fail to select," and "non-select" all of which are replaced with the term "FOS." This is the Navy acronym most often used for a failed officer selection. For each different survey question, we save the preprocessed comments

with unique identification codes as a corpus of "documents," where a document is the text of a single comment.

2. Candidate Tokens

Consecutive-word phrases, referred to as *n*-grams, are any *n* consecutive words found in a comment. Chuang et al. (2012b) demonstrate that there is little added benefit to using *n*-grams of more than three words. As a result, for each comment, we extract all 1- to 3-grams that do not cross any type of punctuation boundary for each comment; we define this set of words and phrases as the set of candidate tokens for that comment.

3. Candidate Token Score

Once constructed, a comment's candidate tokens consisting of the unigram, bigrams, and trigrams are evaluated to determine which has potential to best represent that comment. Chuang et al. (2012b) demonstrate that statistical and linguistic elements of text can be used to determine a descriptive label for larger documents, such as journal articles and dissertations. We show that similar concepts can be used on smaller text comments that are approximately 1000 characters or less. Chuang et al. (2012b) assign candidate token scores (CTS) to each comment's candidate tokens. The token with the largest score among candidate tokens for a particular comment is taken to be the label for that comment. This CTS is a linear function of variables computed for each candidate token to describe a token's potential to describe its comment. There are two types of variables, token specific and comment specific, both types and how we determine their coefficients are described in this section.

a. Token Specific Variables

Token specific variables are unique to each candidate token and independent of the comments associated with the token. These variables are calculated once for each unique candidate token in the corpus and are then factored into the final CTS computation. The four token specific variables are: TS, the number of words in the token known as the token size; TT, an indicator of whether the token is a technical term or not; PTT, an indicator of whether a token is a partial technical term or not; RC, a score to

measure the "commonness" of the token in a reference corpus. These are described in detail in this section.

(1) Token Size

Our candidate tokens are limited to 1- to 3-grams of consecutive words. This creates many tokens that begin with the same word and tokens that contain substrings of other tokens. Although not used in previous research, we treat the token size as a categorical variable to capture the significance that multi-word tokens play in describing a comment.

(2) Technical Terms and Partial Technical Terms

We first determine whether each candidate token is a technical or a partial technical term. Language using technical terms is more descriptive and technical terms are often better to represent larger documents or summaries. Technical terminology has no universally accepted definition. Justeson and Katz (1995) characterize it as words or phrases that have a widespread meaning on their own, but a more specific and accepted meaning when referenced to a specific subset. For text analytic purposes, technical terms are defined as a multi-word phrase that meets a specific pattern; it begins with either an adjective or noun, strings together adjectives, nouns, or prepositions in the middle, and ends in a noun. The only exception is to exclude determiners as adjectives. Further research by Chuang et al. (2012b) suggests that cardinal numbers are useful as a final word of a token since years, version, or iterative numbers may be good descriptions such as *Windows 10* or *DDG 80*. In addition to technical terms, Chuang et al. (2012b) differentiates between technical terms and compound technical terms, where he allows the word "of" as a middle term, such as "*War of 1812*." Because we have short comments with few words and to simplify our model, we combine technical and compound technical terms into a single category that we call technical terms and include the tokens with the word "of" as a middle term in our definition. The variable TT takes value 1 if a candidate token is a technical term and 0 otherwise. Chuang et al. (2012b) further defines partial technical terms that they define as tokens that match a substring of a technical term. For example, the single term "map" is not a technical term since technical terms are

multi-word tokens. If the token "route map" is defined in a reference corpus as a technical term, "map" is a substring of "route map" and meets the definition of partial technical term. We identify whether each candidate token is partial technical term or not. Similarly to TT, we define the variable PTT to take value 1 if a candidate token is a partial technical term and 0 otherwise. To ensure that the definition of PTT is consistent, say across time or no matter what subset of survey responses are analyzed, the candidate tokens used to define PTT are extracted from a reference corpus rather than the comment corpus. We describe such a reference corpus in the next section.

(3) Reference Commonness

Chuang (2013) shows that a measure of how common a token is across a reference corpus or "the normalized term frequency [in the reference corpus] relative to the most frequent *n*-gram [in the reference corpus]" is a factor that can be used to help determine which tokens make the best descriptors. The reference corpus can be the document that the token comes from, the entire corpus of documents being labeled, or an entirely separate corpus created from general web scraping. For any reference corpus, Chuang (2013) argues that the best descriptor tokens come from the middle of the reference corpus term frequency distribution, with extremely frequent and infrequent tokens being less likely to provide good descriptive labels. However, none of the generic corpora suggested by Chuang (2013) appear to be useful as reference corpora for short, Navy specific comments. As a result, we construct the variable reference commonness (RC) using Navy documentation. This approach, selecting a reference corpus with terminology, acronym, and language use that is consistent with the corpus of short-text comments, is easily adaptable to any context where relevant references are available.

Based on the survey topic, there are Navy reference manuals and targeted web pages that outline Navy policy and programs related to most topics. For example, Navy Retention Survey comments include many topics that refer to programs and policies related to personnel. This directly relates to the Department of the Navy's Naval Military Personnel Manual (MILPERSMAN) (2002), which we use as a reference corpus for Navy Retention Survey comments.

We stem the reference corpus to account for occurrences of different tense words with the same meaning and then we calculate the RC for each 1- to 3-gram token contained in the reference corpus by dividing the log of the token frequency in the reference corpus by the log of token frequency of the most frequent token of the same token size. For example, the most frequent 2-gram in the MILPERSMAN is "of the" with a frequency of 5464. The token "duti station" has a frequency of 450. The RC for "duti station" is

$$\frac{\log(450)}{\log(5464)} = 0.71,$$

where RC=0 for a token that does not appear in the reference corpus. The RC is then assigned to a six-level categorical variable that is corresponding to following sets: {0}, (0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], and (0.8, 1.0].

b. Comment Specific Variables

This section discusses variables required for the CTS that are unique to each comment. For surveys with many comments, the processes to compute these variables need to be succinct and efficient. The three comment specific variables are: Freq, the frequency of the token in each comment; RFO, a measure of the first occurrence of token relative to a token of the same frequency; FH, an indication of whether a token is contained in the first half of a comment or not.

(1) Frequency

We use the comment corpus and our candidate tokens to construct a document term matrix (DTM) with one row per document and one column per candidate token that stores each candidate token frequency count by document. Based on a corpus of large documents, Chuang et al. (2012b) demonstrate that tokens that appear more frequently in a document are often more important than less frequent tokens. In our short documents, important tokens may appear only once. As a result, this variable does not have the effect for shorter documents that it has for larger documents. The log of the candidate token frequency is used as the variable for computing the CTS.

(2) Positional Elements

Chuang et al. (2012b) show that the position of a token in reference to the beginning of a document and the length of the document is significant in selecting descriptive tokens. Tokens first introduced closer to the beginning of a document are more important, but the significance is reduced if the token occurs too frequently later in the document. As a result, Chuang et al. (2012b) defines the *absolute first occurrence* (AFO) of a token, as a normalized measure of the location of a token's first appearance in a document with 0 representing the first token in the document and 1 representing the last. For tokens that contain 2- or 3-grams, AFO is calculated using the normalized position of the first word in the phrase and the total number of words, counting the *n*-grams as a single "word" for this calculation. This creates a distinction so that a unigram, bigram, and trigram that all begin at the same location do not all have the same absolute first position except for the tokens that begin the comment.

From the AFO, the *relative first occurrence* (RFO) is defined. It "measures how likely a term is to initially appear earlier than a randomly-sampled phrase of the same frequency" (Chuang, 2013). Let *k* be the frequency of a token in the document, then

$$RFO = (1 - AFO)^k. \tag{2.1}$$

Chuang et al. (2012b) indicates that tokens in the *first sentence (FS)* are more important and are often better descriptors than tokens later in a document. However, short comments with 1000-character limits are a few words, a single sentence, or a short paragraph at most. To make a positional comparison, our method needs a variable with a similar ratio as a sentence length to paragraph length ratio. We define two new variables the *first quarter (FQ)* and *first half (FH)* as binary variables that correspond to whether a candidate token is fully contained in the first quarter or the first half of a comment. We find that these variables add more value than the first sentence indicator variable used by Chuang et al. (2012b), with the FH producing better descriptive labels than the FQ variable. As a result, we utilize the *first half* variable in our CTS computation.

c. Candidate Token Score Calculation

Once all the necessary variables are constructed, the CTS is calculated for each candidate token and comment combination as a linear function of the token and comment specific variables using coefficients estimated using the approach of Chuang et al. (2012b). They determine two sets of coefficients from a two diverse data sets. The first is a corpus of 144 dissertations published at Stanford University between 1993 and 2008 with various topics across six departments. Sixty-nine student volunteers from these departments select keywords to describe the dissertations. Token and comment specific variables are computed for each keyword. Researchers exclude keywords for which values for token or comment specific variable cannot be determined. This includes keywords that are not in the dissertations, for which positional variables, such as FS, cannot be computed, and phrases that are longer than five words, since for Chuang et al. (2012b) the web commonness only includes up to 5-grams. The result is 2,882 usable observations. The second data set consists of 244 scientific articles with four subdisciplines in the Association for Computing Machinery (ACM) Digital Library as selected for the Semantic Evaluation 2010 contest (Kim, Medelyan, Kan, and Baldwin, 2010). Keywords provided with the dataset are author-assigned, reader-assigned, and a combination of author and reader-assigned and are used as expert observations for this dataset. The usable observations of the two datasets are taken to be positive responses indicating that the corresponding keyword is an appropriate label for the document. Topic and comment specific variables are also computed for each of these keywords. Ten additional randomly selected tokens are generated from the corpus for each document to use as negative responses and assigned a weight of 0.1. A logistic regression model is used with the variables and weighting to estimate a set of regression coefficients for each dataset.

Using the essence of this work, we randomly select fifty comments that include at least five words from each of four questions across three surveys to construct our dataset. The comment that clarifies the *Other* reason sailors are leaving the Navy and the comment "What can be done to encourage you to remain in the Navy on active duty when you are next required to make a stay/leave decision?" are both from the Navy

Retention Survey. The Female Dress Uniform & Cover Survey conducted by OPNAV N1 in 2017 includes the questions, "What do you like about the SDB slacks?" and "What changes do you desire to the female SDB coat?" (MAX.gov, 2017). These four questions cover the two topics of manpower and uniforms and both have references including the MILPERSMAN and Navy Uniform Regulations (Department of the Navy, 2011) that are used for the reference copra. We read the comments to determine the primary 1- to 3-gram label for each comment and store them as the *expert label*. We create all 1- to 3-gram tokens for each comment, excluding the *expert label*, and randomly select ten tokens to use as negative responses. This produces a data set of 2,200 comments with labels. We determine the variable values for each label comment pair and fit a logistic regression to estimate our own regression coefficients. Our coefficients are shown in Table 1.

Table 1. Regression Coefficients for Candidate Token Score Calculation

Model Variable	Coefficient Estimate	Standard Error
(Intercept)	-2.5217800 ***	0.6249
TS = 2	-1.1246629 **	0.4086
TS = 3	-1.2805272 **	0.4512
TT	3.2928379 ***	0.5675
PTT	-1.0478745 *	0.4384
RC ϵ (0%, 20%]	-0.2783750	0.5632
RC ϵ (20%,40%]	-0.8293879 •	0.4572
RC ϵ (40%, 60%]	-0.5307969	0.4456
RC ϵ (60%, 80%]	-0.1845462	0.5033
RC ϵ (80%, 100%]	-2.9243750 *	1.1700
Log(Freq)	0.5744200	0.8267
RFO	3.8014855 ***	0.8273
FH	0.3303841	0.5082

Statistical significance = ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, •: $p < 0.1$

4. Comment Label

The candidate tokens for each comment are scored using the coefficients from Table 1. The candidate token with the maximum CTS among candidate tokens for a comment is assumed to be the label for that comment.

B. GROUP COMMENTS INTO SIMILAR BINS

The comment labels are used with more traditional visual text mining methods to identify primary topics among the comments. Once labels are determined, we use them to choose words or phrases, known as keywords, to assign to each observable theme from the comments. We compare keywords to the comment labels and original text to group comments into topic bins.

1. Preprocess Labels and Document Term Matrix

The comment bin assignment process begins with constructing a corpus of the comment labels where each document of this corpus corresponds to a label. Here, *n*-gram labels are rendered into documents of *n* words. Words in the corpus are stemmed. This allows words such as "assigning," "assignment," and "assignments" to be converted to their root "assign" so that a frequency of the idea can be captured regardless of the tense or inflection used in the label. A DTM is formed from the stemmed corpus.

2. Determine Topic Bins

Although traditional text analytic methods are not useful on short survey text comments, using these methods on comment labels is beneficial to determining topics. LDA is a common topic modeling technique that identifies the latent topics of a corpus of documents while allowing multiple topics for each document. It estimates the distribution of words in each topic and the distribution of topics for each document (Silge & Robinson, 2017). Our method uses an LDA model to estimate these distributions.

We fit the LDA model to determine the number of topics using the log-likelihood as a function of the number of topics. Estimates of the topic distribution $\{P(T|d)\}$ for

each document *d* in the corpus and estimates of the word (token) distributions $\{P(w|T)\}$ for each topic *T* are extracted from the LDA fit.

Zhai and Massung (2016) contains details of these computations, but we reproduce them here for completeness. Let N_d be the number of tokens in a document *d* and $N = \sum_d N_d$ be the total number of tokens in the corpus. Then the expected number of tokens in the corpus associated with a topic *T*, N_T, is given by

$$N_T = \sum_d N_d P(T|d), \tag{3.1}$$

and the proportion of such words in each topic or the topic distribution across words in the corpus $\{p(T)\}$ is given by the ratio N_T / N. The estimate that a given token (word) comes from topic *T*, $p(T|w)$, is found using Bayes rule,

$$p(T|w) = \frac{p(w|T)p(w)}{P(T)}. \tag{3.2}$$

From $\{P(T|w)\}$ for each token *w* and $\{P(T)\}$, we find the distinctiveness defined by Chuang et al. (2012a) of each token *w* using the Kullback–Leibler divergence (Kullback and Leibler, 1951), which measures how much $\{P(T|w)\}$ diverges from $\{P(T)\}$,

$$distinctiveness(w) = \sum_T P(T|w)\, log\left(\frac{P(T|w)}{P(T)}\right). \tag{3.3}$$

Words for which $\{P(T|w)\}$ is "close" to $\{P(T)\}$ have distributions close to zero and are not words that help identify distinct topics.

Chuang et al. (2012a) introduce *saliency* as a measure used to find relevant but not overly frequent words in topics, which we find useful for identifying topic bins. It is the product of the probability of selecting a term *w* from the corpus of words, $P(w)$, and the distinctiveness. Saliency of word is defined as:

$$saliency(w) = P(w) * distinctiveness\ (w). \tag{3.4}$$

Saliency provides another measure in addition to frequency that can be easily visualized for topic identification to ensure the major topics are identified.

Once frequency and saliency are computed for each word in the label corpus, we construct a correlation plot of the DTM twice using the most frequent and most salient words. We adjust the correlation threshold and number of terms until we display a plot that illustrates key topics. From the two plots and using background knowledge on the survey question, we determine primary topics that become our *topic bins*.

3. Create Topic Bin Key

With defined topic bins, we (the analysts) continue to review correlation plots to determine keywords that define each topic based on the comment question. Correlation plots display a visual representation of tokens from a DTM based on a user-defined term list. Connections are displayed between tokens that have a correlation that exceeds the defined correlation threshold. The thickness of the arc is proportioned to the strength of the correlation, with thicker lines indicating a stronger correlation.

In addition, we remove stop words from the label corpus and then display the corpus in a word cloud where the size of the word is proportional to its frequency. This allows ease of viewing more frequent words. This visual display supports the identification of keywords that belong in each topic bin and helps to determine if any high frequent terms are unassigned and require an additional bin. This important step allows background knowledge of subject matter experts to define logical and meaningful bins. The keywords and associated topic bins are stored in the *topic bin key*.

4. Assign Comments to Topic Bins

Assigning comments to topics is an iterative procedure. First, using the topic bin key as a lookup reference, comment labels are compared to the keywords and assigned to the corresponding topic if there is a match. Since the comment labels are selected to be the most descriptive tokens from the comments, these are used as the first line of automatic assignment. However, some comments have labels that do not match keywords but still fit in a bin based on the other words in the comment. In cases where a label is not binned, we search for matches between the comment text and the topic keywords, and assign matches to corresponding topics. In the final step, we review a word cloud of the comments that are not assigned to any bin. If prevalent keywords are found to indicate

additional topics or keywords, the topic bin key is updated and the assignment algorithm is repeated on all comment labels. This process of topic assignment and review repeats until there are no unassigned comments, the remaining comments cannot be assigned to a topic, or we reach an acceptable threshold of unassigned comments. Once this level is reached, we label these comments as *Other* and bin them together.

C. COMMENT ANALYSIS APPLICATION

In this section, we describe how we implement our methods with examples of each step. The first section describes labeling the comments and uses a comment from the Milestone Survey for demonstration. The second section demonstrates our iterative process for grouping comments into similar bins.

1. Label Comments

A sample comment will be used from the Milestone Survey to demonstrate how we obtain labels for comments.

Example Comment:

Duty stations (& career assignment) are major factors in considering to stay Navy.

a. Preprocess Data

Comments are imported into the statistical computing environment R (R Core Team, 2017) as a data frame and require minimal processing prior to applying text analysis techniques. Required steps include filtering out skipped comments that are represented with an "S," converting text to all lower case, replacing standard contractions with their non-contraction equivalent, removing non-sentence defining punctuation, and replacing similar words or phrases and Navy specific acronyms with a single representation. Appendix D includes the list of contraction and word substitutions. Unnecessary punctuation is also removed. Once the data is preprocessed, the comment and ID are saved into a corpus

Preprocessed Comment:

duty stations and career assignment are major factors in considering to stay navy.

b. Candidate Tokens

The **RWeka** (Hornik, Buchta, & Zeileis, 2009) function `NGramTokenizer()` is a tokenizer in R that creates *n*-grams based on the minimum and maximum number of tokens entered by the user. We use the `DocumentTermMatrix()` function from the **tm** package (Feinerer & Hornik, 2017) with `NGramTokenizer()` to construct 1- to 3-gram tokens and store their document frequency count in a DTM. The resulting 1- to 3-grams are defined as the candidate tokens to describe each comment. Table 2 shows the selection of candidate tokens for the example comment.

Table 2. Candidate Tokens

and	factors
and career	factors in
and career assignment	factors in considering
are	in
are major	in considering
are major factors	in considering to
assignment	major
assignment are	major factors
assignment are major	major factors in
career	navy
career assignment	stations
career assignment are	stations and
considering	stations and career
considering to	stay
considering to stay	stay navy
duty	to
duty stations	to stay
duty stations and	to stay navy

c. Variable Calculations

This section explains the R process for each variable calculation as described in Chapter II, Section A, Subsection 3c.

(1) Token Size

We use regular expressions to determine the number of words contained in each token and assign this as the token size. Table 3 displays the TS for each variable.

Table 3. Token Size

and	1	factors	1
and career	2	factors in	2
and career assignment	3	factors in considering	3
are	1	in	1
are major	2	in considering	2
are major factors	3	in considering to	3
assignment	1	major	1
assignment are	2	major factors	2
assignment are major	3	major factors in	3
career	1	navy	1
career assignment	2	stations	1
career assignment are	3	stations and	2
considering	1	stations and career	3
considering to	2	stay	1
considering to stay	3	stay navy	2
duty	1	to	1
duty stations	2	to stay	2
duty stations and	3	to stay navy	3

(2) Technical and Partial Technical Terms

The first step in identifying technical and partial technical terms is to identify parts of speech (POS) elements for the words of candidate tokens using the Stanford POS Tagger and the **openNLP** wrapper package (Hornik, 2016). Since candidate tokens are between one and three words, they all have at most one word that represent the first, middle, and last word of a technical term. Using R, candidate tokens are assessed to see if they meet the constraints defining them as a technical term. If so, TT is assigned value 1 and 0 otherwise. Candidate tokens are also reviewed to see if they are a substring of a technical term in the comment or the reference corpus and assigned a value 1 is they are a substring and 0 otherwise.

(3) Reference Commonness

The MILPERSMAN (Department of the Navy, 2002) is used as the reference corpus and is readily available in a Portable Document Format (PDF). The PDF file is imported into R using `pdf_text()` in the **pdftools** package (Ooms, 2017) and is stored as a list, where each page of the PDF is a separate element. The first three and last line of each page containing the header and footer are removed. The document is preprocessed to convert all words to lowercase, convert contractions, remove all punctuation, and finally stemmed. All pages are converted to a single string and the reference corpus is constructed. The `DocumentTermMatrix()` function is used to produce a DTM with unigrams, bigrams, and trigrams. The RC is computed for each set of *n*-grams based on their frequency in reference to the most frequent *n*-gram of the same size. The candidate tokens are stemmed and matched to tokens from the reference corpus. Candidate tokens are assigned RC=0 if they do not match, and inherit RC from the reference token if they do. The six-level categorical variable used to compute CTS is then constructed from RC.

(4) Frequency

The frequency of candidate tokens for each document is extracted from the comment corpus DTM.

(5) First Half

Using the `gregexpr()` function, each comment in the corpus is divided using an alphanumeric regular expression to determine the total number of words in the comment. The total number is divided by 2 and rounded up to the nearest whole number to determine the cutoff position for the first half of the comment. Each comment is truncated using the `strsplit()` function to include only the first half of the comment and stored separately. Candidate tokens are compared to the truncated comment to determine if they appear in the first half using an exact regular expression match. The first half variable, FH, is the indicator function with value 1 if the entire token is in the first half and 0 otherwise.

Example First Half:

duty stations and career assignment are major

d. Candidate Token Score Calculation and Label

The CTS is calculated using the sum product of the regression coefficients from Table 1 and the corresponding variable values. In Table 4, we display the score calculations for the example comment. The three-level categorical variables TS and the six-level categorical variable corresponding to RC are replaced with a single column labeled TSC and RCC, respectively. The values in these columns are each token's TS and RC contribution to CTS. For all other variables, the Table 1 coefficients are given in the first row of Table 4. The candidate token with the maximum CTS, "duty stations" in our example, is assigned to be the label for the original comment.

Table 4. Variable Summary with Final Candidate Token Score

Coefficient	**0.57**	**1**	**1**	**3.80**	**0.33**	**3.29**	**-1.05**		
Token	log(Freq)	RCC	TSC	RFO	FH	TT	PTT	CTS	Rank
and	0	-2.92	0	0.83	1	0	0	-1.95	20
and career	0	-0.83	-1.12	0.82	1	0	0	-1.04	12
and career assignment	0	0	-1.28	0.8	1	0	0	-0.43	8
are	0	-0.18	0	0.58	1	0	1	-1.21	14
are major	0	-0.28	-1.12	0.55	1	0	0	-1.52	16
are major factors	0	0	-1.28	0.5	0	0	0	-1.9	19
assignment	0	-0.18	0	0.67	1	0	1	-0.89	11
assignment are	0	-0.53	-1.12	0.64	1	0	0	-1.43	15
assignment are major	0	0	-1.28	0.6	1	0	0	-1.19	13
career	0	-0.18	0	0.75	1	0	1	-0.57	9
career assignment	0	-0.83	-1.12	0.73	1	1	1	0.86	2
career assignment are	0	0	-1.28	0.7	1	0	0	-0.81	10
considering	0	-0.18	0	0.25	0	0	0	-1.76	17
considering to	0	-0.53	-1.12	0.18	0	0	0	-3.49	31
considering to stay	0	0	-1.28	0.1	0	0	0	-3.42	30
duty	0	-2.92	0	1	1	0	1	-2.36	21
duty stations	0	-0.18	-1.12	1	1	1	1	2.55	1
duty stations and	0	-0.53	-1.28	1	1	0	0	-0.2	6
factors	0	-0.53	0	0.42	0	0	1	-2.52	22
factors in	0	-0.83	-1.12	0.36	0	0	0	-3.09	29
factors in considering	0	0	-1.28	0.3	0	0	0	-2.66	25
in	0	-2.92	0	0.33	0	0	0	-4.18	35
in considering	0	0	-1.12	0.27	0	0	0	-2.61	24
in considering to	0	0	-1.28	0.2	0	0	0	-3.04	28
major	0	-0.53	0	0.5	1	0	1	-1.87	18
major factors	0	-0.28	-1.12	0.45	0	1	1	0.05	4
major factors in	0	-0.28	-1.28	0.4	0	0	0	-2.56	23
navy	0	-0.18	0	0	0	0	0	-2.71	26
stations	0	-0.18	0	0.92	1	0	1	0.06	3
stations and	0	-0.53	-1.12	0.91	1	0	0	-0.39	7
stations and career	0	0	-1.28	0.9	1	0	0	-0.05	5
stay	0	-0.83	0	0.08	0	0	0	-3.03	27
stay navy	0	0	-1.12	0	0	0	0	-3.65	33
to	0	-2.92	0	0.17	0	0	1	-5.86	36
to stay	0	-0.28	-1.12	0.09	0	0	0	-3.58	32
to stay navy	0	0	-1.28	0	0	0	0	-3.8	34

2. Group Comments into Similar Bins

Once the comment labels are assigned, we store them as a separate corpus, use them to determine topic bins, and later assign them to a bin.

a. Preprocess Labels and Construct Corresponding DTM

Preprocessing the labels begins with a general review of all comments for common terms indicating that no valuable information is included. Comments beginning with the words "none," "nothing," "na," "not applicable," "no comment" or containing fewer than three characters are automatically categorized and binned as "no comment." From the remaining comments, associated labels are extracted for analysis. A corpus is constructed with the comment labels using the **tm** function `VectorSource()`. The corpus is stemmed and the DTM is constructed also using functions from the **tm** package. Empty rows are removed from the DTM and the `LDA()` function from the **topicmodels** package (Grün & Hornik, 2011) is used to train an LDA model with two to fifteen topics. The "best" number of topics can be determined by locating the "knee" in a log-likelihood plot. Alternatively, we automate this process by using the log-likelihood values and taking the third difference (Chen et al., 2017). We determine the index of the maximum absolute value of the third difference and add two to the position to account for the differencing. The result will contain 3 to 14 topics. The LDA model is then fit to the number of topics found. We estimate the distinctiveness and saliency of each token *w* using (3.3) and (3.4) respectively.

The most frequent and the most salient terms are displayed in correlation plots by using the `plot()` method for DTMs from the tm package. Arguments for this function are the labeled corpus DTM vector of the most frequent (or most salient) terms. In these plots, the thickness of the arcs between terms proportional to the correlation where no arc is plotted if the correlation is less than the correlation threshold that we set. Examples, displayed in Figures 1 and 2, are reviewed to determine the topics bins.

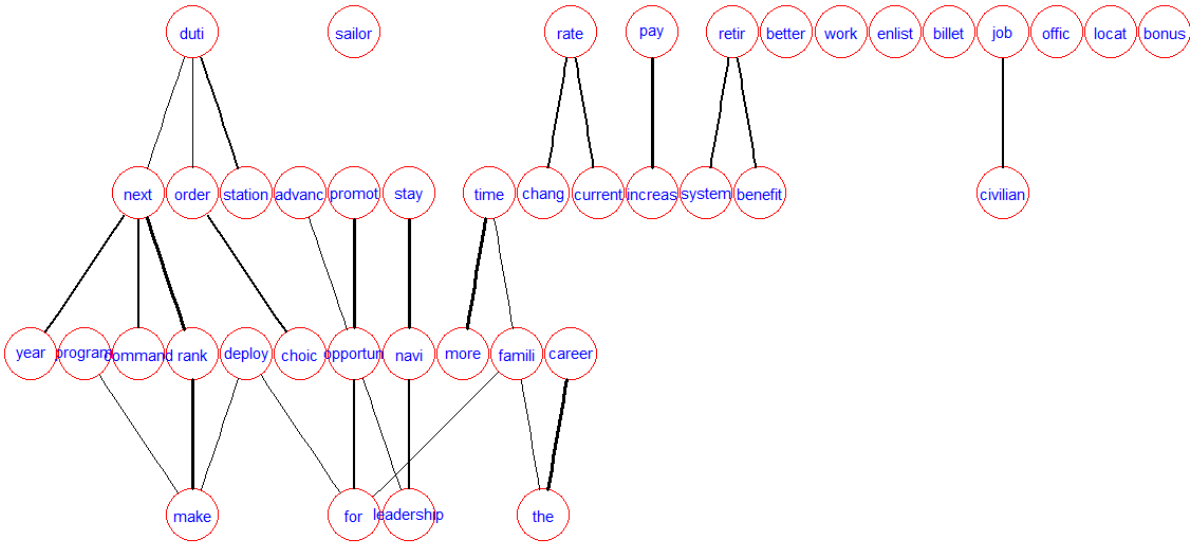


Figure 1. Correlation Plot of the Most Frequent Terms from the Labels

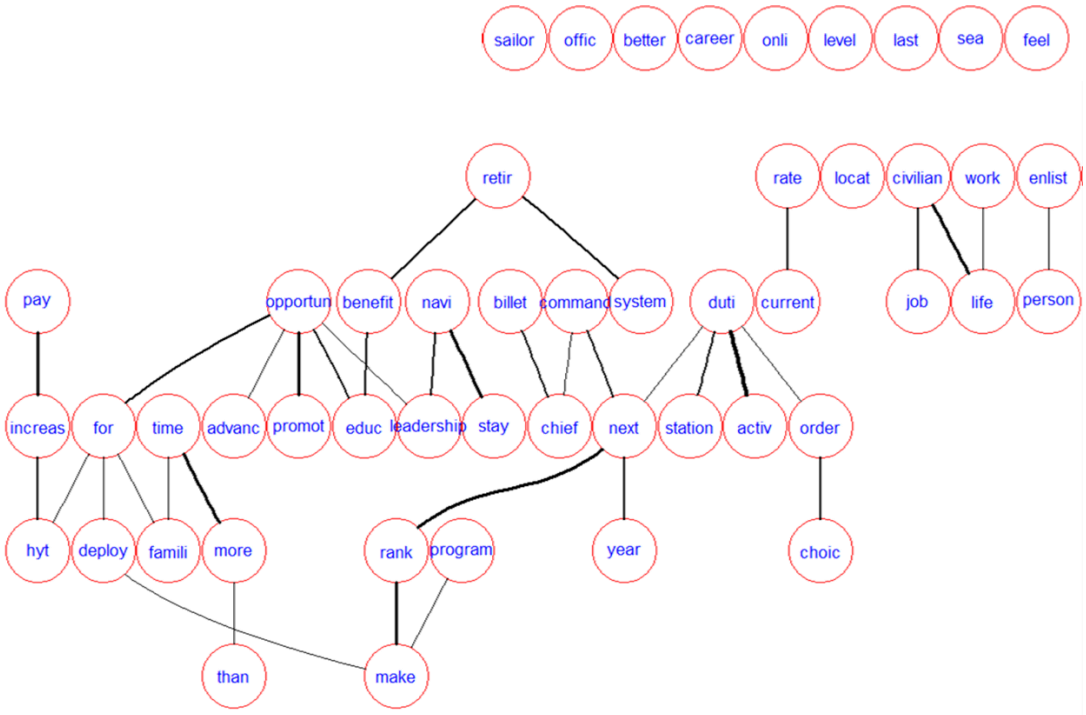


Figure 2. Correlation Plot of the Most Salient Terms from the Labels

Example prevalent topics from the correlation plot based on frequencies in Figure 1, include "choic order," "make rank," "next duty station," "advanc opportun," "promot opportun," "more time family," "chang rate," "increas pay," "retire system," "bonus," and "locat." Reviewing the saliency plot uncovers additional Navy topics such as "educ opportun," "increase hyt," and "work life." These topics are used as a basis to construct the topic bins found in the topic bin key (Appendix E).

b. Construct Topic Bin Key

Next, we remove stop words from the label corpus and create a word cloud from the corpus, as seen in Figure 3. The word cloud highlights frequently use terms, but the less frequent terms are also available to assist in creating a key that is used as look up table to assign comments to each bin.

Figure 3. Word Cloud of Stemmed Label Corpus

The topic bin phrases “next duti station” and “locat” appear to indicate that respondents would be encouraged to stay on active duty if they can be in a specific location or choose their next duty station. Using this as an example topic bin, we review the word cloud for additional related keywords. Potential words include, “geograph” and “area.” Using an understanding of this topic, additional keywords are added based on prior experience with the topic including keywords such as “homestead,” “choose homeport,” “closer” to a location or family, “not transfer” from current location, stay in one “region.” From this analysis, a topic bin key, containing all topics and their associated keywords, is constructed. The complete topic bin key used for the encouragement to stay analysis can be found in Appendix E.

c. Assign Comments to Topic Bins

The topic bin key is first compared to the labels. Using regular expressions to allow for partial matches, labels are searched for each keyword from the topic bin key and the positions of the matches are saved and compared. The match that appears earliest in the label is considered the primary topic and the label is assigned to that keyword’s corresponding topic bin. For labels that do not contain any keyword matches, the comments are reviewed to determine if they contain keywords. Matches are assigned to corresponding topic bins. A table of frequency counts of the topic bins is constructed to determine if binning the remaining comments as *Other* would be acceptable. If *Other* is the largest category, or very close to the top, a word cloud of the words in the remaining comments is reviewed to construct additional bins or keywords and the process continues until an acceptable number of comments are assigned to primary topics.

THIS PAGE INTENTIONALLY LEFT BLANK

III. ALL NAVY APPLICATION TO RETENTION SURVEYS

In this chapter, we illustrate our method on the Career Viewpoint Retention survey, constructed to collect retention opinions from sailors as part of the Career Viewpoint strategy. Career Viewpoint is a concept that was initiated in 2013 to improve survey administration in the Navy. The concept developed into the construction of the Career Viewpoint Surveys and Studies (CVSS) application and the Navy Retention Survey. In the next section, we provide a detailed background of the Career Viewpoint strategy, including the survey application, the Navy Retention Survey, and the available reports of the results. In the second section, we discuss DARTS, the tool that provides additional demographic filtering capability and the inclusion of our binned comments. The final section provides results from the Retention Survey and demonstrates implementation of DARTS.

A. CAREER VIEWPOINT BACKGROUND

The Bureau of Naval Personnel (BUPERS), Military Community Management (BUPERS-3) with the support of OPNAV Military Personnel Plans and Policy Division (N13), developed the Navy Retention Survey in 2014. The 2010 elimination of the previous retention survey sparked this development to produce a survey that was less costly to administer with a higher response rate. Along with the Navy Retention Survey, the CVSS application for disseminating three versions of the Navy Retention Survey was developed and adopted.

1. Career Viewpoint Surveys and Studies (CVSS)

CVSS is a web-based survey and analysis application that uses the Navy's personnel management infrastructure of the Navy Standard Integrated Personnel System (NSIPS). It targets sailors directly based on demographics and military specific factors that are recorded in their Electronic Service Record (ESR). This application allows the survey administration process to stay contained within a secure environment while providing confidentiality of the responses and archival storage (Lockheed Martin, 2013).

CVSS was developed with the intention of disseminating the Navy Retention Survey, but the system proves beneficial for surveys that require short turnaround times and increased participation. The Enlisted Women in Submarines Survey was a 22-question survey that polled enlisted females on their interest in the submarine community. It was the first survey to use CVSS and served as a proof of concept for rapid turnaround. Once the survey was officially approved, it was entered into CVSS, tested by key stakeholders using CVSS, revised, was ready for the fleet, and disseminated to 50,449 AC/FTS enlisted females in one month. The survey utilized automatic emails sent to females with valid email addresses in NSIPS, a Navy-specific administrative message (NAVADMIN), and Navy Times article to request participation. It remained open 31 March – 30 April 2014 and exceeded typical Navy survey participation rate of 20% with 26% participation with a one-month deployment instead of the recommended and typical two-month deployment (Career Viewpoint Surveys and Studies, n.d.).

Because this application is housed within the Navy personnel system used to manage all personnel records, CVSS can extract over 100 demographics and military specific details about a survey respondent when a survey request is made. See Appendix A for a list of available elements. This eliminates the need to have a member answer these questions, which reduces the length of the survey. It also allows the survey questions to be filtered based on a member's status. For example, many questions refer to a spouse or children. These questions are tagged so that they only display for members who show the correct dependent code within their member record. Additionally, capturing this data from the record more accurately reflects member status, since a member does not have the option to misrepresent their personal information. Lastly, the survey and its data are archived so that they can be used to answer future questions that were not considered when the survey was created.

2. Career Viewpoint Retention Survey

As outlined in a memorandum approved by Director, N13 in 2013, there are two approved versions to the Career Viewpoint Retention Survey. The first is the Career Viewpoint Exit Survey (Active Component (AC)/Full Time Support (FTS)). This survey

targets sailors who have an indication in their record that they are leaving active duty service. The second is the Career Viewpoint Milestone Survey (AC/FTS), which targets members with service time remaining that are eligible for retention. In addition to the approved version, the Career Viewpoint Reserve Survey was developed and is available for future dissemination.

a. Career Viewpoint Exit Survey

The Career Viewpoint Exit Survey (AC/FTS) has the primary purpose of determining why sailors leave active duty in the Navy. Members are sent an email to their official NSIPS email address requesting their participation in the survey six months prior to their Estimated Date of Loss to the Navy (EDLN) or when an enlisted member has a Career Waypoints (C-WAY) status indicating that they are leaving the Navy. The survey can be requested by a member's career counselor within their NSIPS Career Information Management System (CIMS) account. A member can also self-request the survey within their ESR by following the menu path "Employee Self Service," "Electronic Service Record," "Tasks," "Survey Requests," and selecting Survey Request ID: "1000000024."

b. Career Viewpoint Milestone Survey

The second version is the Career Viewpoint Milestone Survey (AC/FTS), which targets members with service time remaining on active service that are in a window to make a stay or leave decision. The survey provides measures that show indications if members plan to stay on active duty or leave for various reasons. The survey is available to officers 15 months prior to their minimum service requirement (MSR) or projected rotation date (PRD). This is approximately three months prior to when a member must either negotiate orders for another tour or officially indicate that they intend to resign their commission. Enlisted members receive the survey 18 months prior to their Soft End of Active Obligated Service (SEAOS), or 5 months prior to when the reenlistment request process begins in the C-WAY system. These time frames are set to better ensure that a member can indicate their intentions and opinions about the Navy prior to any request for reenlistment or orders. This provides responses that are less likely to be tainted by the detailing or C-WAY processes and more indicative of a member's tours in the Navy.

c. Career Viewpoint Reserve Survey

The Reserve Retention Survey is the third variation that was created as a part of the Navy Retention Survey. It contains questions that are modified versions of the questions used for the active component. Although this version is available, the reserve component leadership has not approved it and no dissemination of the survey has occurred.

d. Survey Deployment

The Exit and Milestone versions of the survey were deployed 01 July 2014 and have been automatically deployed monthly based on their predefined criteria. They are available to selected respondents for two months. The surveys are comprised of a maximum of 150 questions tailored to the individual taking the survey according to the way they answer the 15 core questions and their demographics as reflected in NSIPS. Most of the questions utilize a seven-point scale representing a sailor's stay or leave tendency toward each question asked. An example is displayed in Table 5.

Table 5. Seven Point Scale Questions

On a sliding scale of 1–7, with 7 being the strongest influence to stay, please indicate if the following factors influence you (contribute to your decision) to stay on active duty, leave active duty, or have no effect on your Navy career intentions.	Leave----- No Effect -------Stay						
	1	2	3	4	5	6	7
Promotion/Advancement opportunities	O	O	O	O	O	O	O
Career assignments (number of options, control over PCS assignments)	O	O	O	O	O	O	O
Command climate (previous and current commands)	O	O	O	O	O	O	O
Work-life balance (operational work demand, sea duty, time away from home)	O	O	O	O	O	O	O

Additionally, there are multiple option questions, multiple option questions with a comment box to clarify a response, and stand-alone comment boxes ranging from 100 to 1000 characters.

e. Displaying Survey Results

Tools available to display the Navy Retention Survey results are limited. Extracting the data from CVSS is difficult for an analyst unfamiliar with the application

and building reports within the analytic environment has a steep learning curve. Because of these factors, the only survey reports available are general summaries that are broken out for officers and enlisted members separately.

(1) PeopleSoft

The results from the Navy Retention Surveys are stored immediately in the PeopleSoft component of NSIPS. This "front end" side can generate a one-page summary report filtered by the survey version, survey expiration date, and command Unit Identification Code (UIC). Personnel with CIMS access within NSIPS can view this report for UIC's assigned to their account. The command career counselor typically has this account access at each command.

This standard report is available separately for the Exit and Milestone Surveys. The report is split into two columns to displays results separately for officers and enlisted members. The available results include participation rates, career intentions, top 5 core stay and leave indicators, top 10 detailed stay and leave indicators, and a policy question summary. Results are only generated if there are 10 or more responses based on the filters to maintain confidentiality. See Appendix B for a sample report.

(2) BusinessObjects

At the beginning of each month, the survey results are updated and compiled in the BusinessObjects (BO) part of CVSS, which is the analytic database side of NSIPS. From each survey requested, five parts are stored within CVSS: (1) survey request details, (2) survey details, (3) respondent's demographic information, (4) respondent's military status details, (5) survey responses. The survey request details are created when the survey is sent and include the predefined criteria used to select the respondent and the survey version information. The survey details include the most recent question and response versions. These are stored to archive changes over time of questions and the choices available for selection to each question. For demographics and military information, a snapshot of over 100 elements is taken from the member's Navy personnel record and is attached to a unique survey request ID when the survey is requested. These elements are updated if a member completes the survey to capture details that may have

changed from the time the survey was requested to the time the survey is taken. The survey responses are saved as a member completes each page of the survey and the survey response storage is finalized when a member completes the survey or the survey expires, with different indicators included for tracking a member's progress.

Included in the BO side of CVSS is a seven-page summary report that is maintained by Space and Naval Warfare (SPAWAR) Systems Center Atlantic. This report includes the same elements as the one-page PeopleSoft report but is displayed with bar and pie charts. This report allows for additional but limited filtering based on zone, duty status, and a roll up of senior UIC's with their subordinate commands as they are assigned within the NSIPS command structure table. The report maintains the 10-response requirement. A sample BO report is available in Appendix C.

In addition to the standard report, the BO environment gives analysts the ability to review or export the raw data and manipulate the data in an ad hoc environment. While complex, ad hoc capabilities include the ability to recreate the SPAWAR controlled BO standard reports, enabling them to be filtered on all demographic and military elements. Additional pre-built reports can be created, such as a demographic summary, to complement the standard report as seen in Figure 4. The advantage of building reports within the BO environment is that the data is updated monthly and the pre-built reports can be recompiled automatically or with little effort from an analyst.

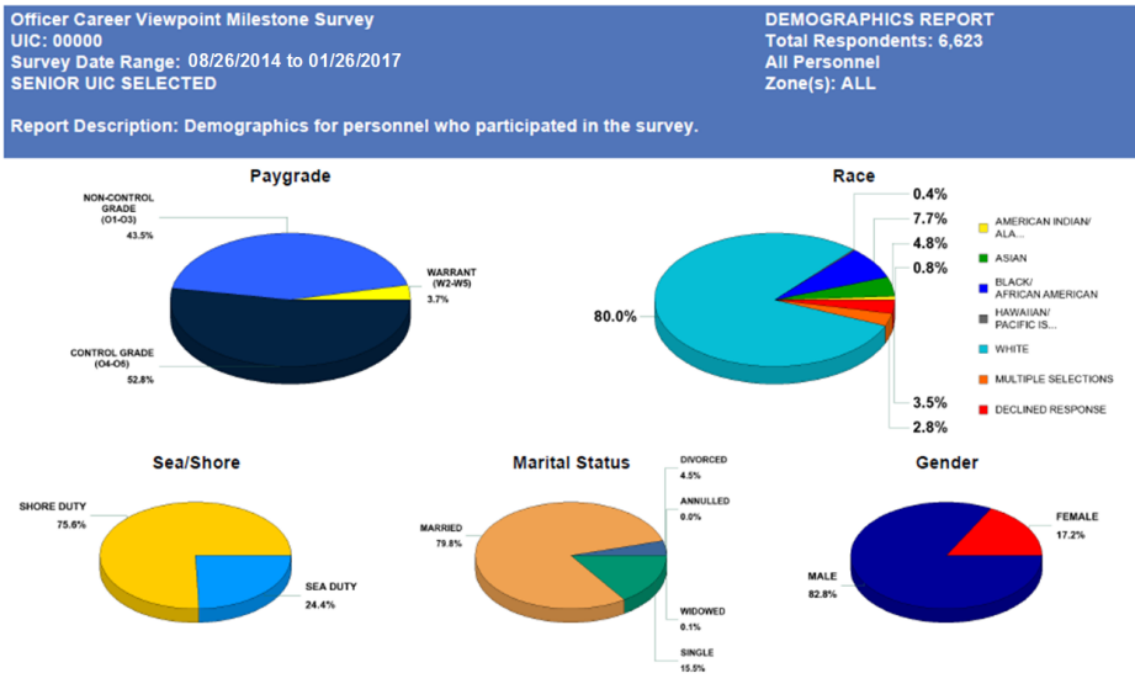


Figure 4. Analyst BO Built Demographic Summary

f. Result Limitations

Access to CVSS is strictly controlled and limited to personnel with a "need to know" requirement who meet proper security clearances. There are approximately five people with access to the application, while only one has a complete understanding of how to efficiently manipulate the ad hoc BO environment. The BO environment is also impractical for text analysis. With the numerous comments from these surveys, this data needs to be extracted and analyzed in another application.

B. DEMOGRAPHIC ANALYSIS OF RESPONSES TOOL FOR SURVEYS (DARTS)

Our comment analysis methodology provides an invaluable resource to bin comments and is enhanced with the use of our Excel based tool DARTS for filtering capabilities. Responses to survey questions often vary depending on service member demographics and experience levels. It is important to extract key groups of respondents to determine the best course of action when making recommendations or decisions based on survey responses.

DARTS is built in Microsoft Excel using Visual Basic for Applications (VBA). It requires, as inputs, the data that contains binned survey comments with corresponding demographic data. The baseline tool is constructed to provide filtering of requested demographic variables by the Navy manpower domain. This capability is especially important for the Diversity and Inclusion branch (OPNAV N1D).

A graphical user interface (GUI) is built for DARTS that provides the user with point and click functionality that easily filters selected measures. The survey date range is particularly important for continuous surveys, such as the Career Viewpoint Retention survey. Additional filters include member type, marital status, race, paygrade band, years of service, and community. Figure 5 displays the baseline GUI for DARTS. It also demonstrates that the corresponding paygrade bands and communities will be displayed based on which member type is selected to account for the differences in officer and enlisted populations. This tool is modifiable based on available demographics and can be updated prior to the completion of a survey. This allows our comment analysis method to be applied and the results added to DARTS immediately upon closing the survey.

DARTS is modified to include a report with survey specific details. This report format is built to be consistent with the Career Viewpoint BO standard reports and is updated with the survey name, report name, and report description. The survey date range and total number of responses are updated automatically when the report is compiled from the GUI to account for filtering. Figure 6 displays a sample report filtered for all Navy officers.

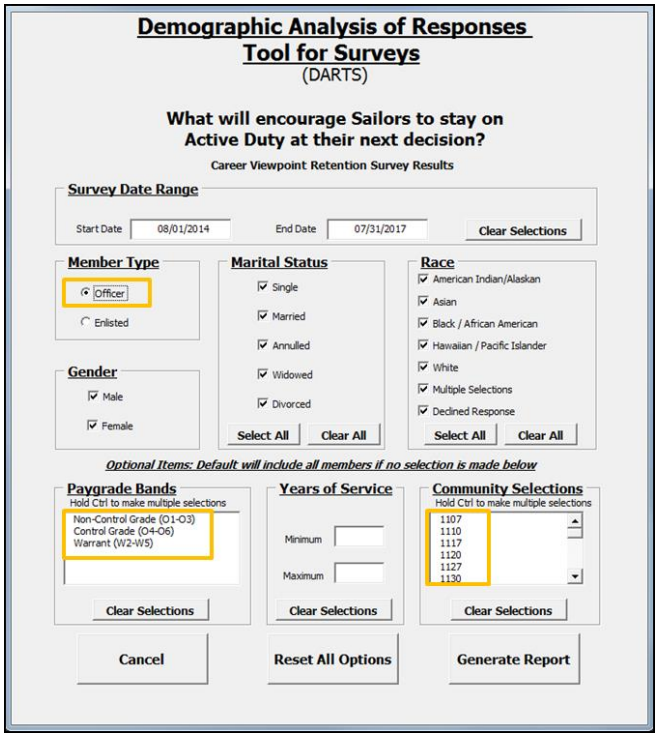


Figure 5. Graphical User Interface for DARTS

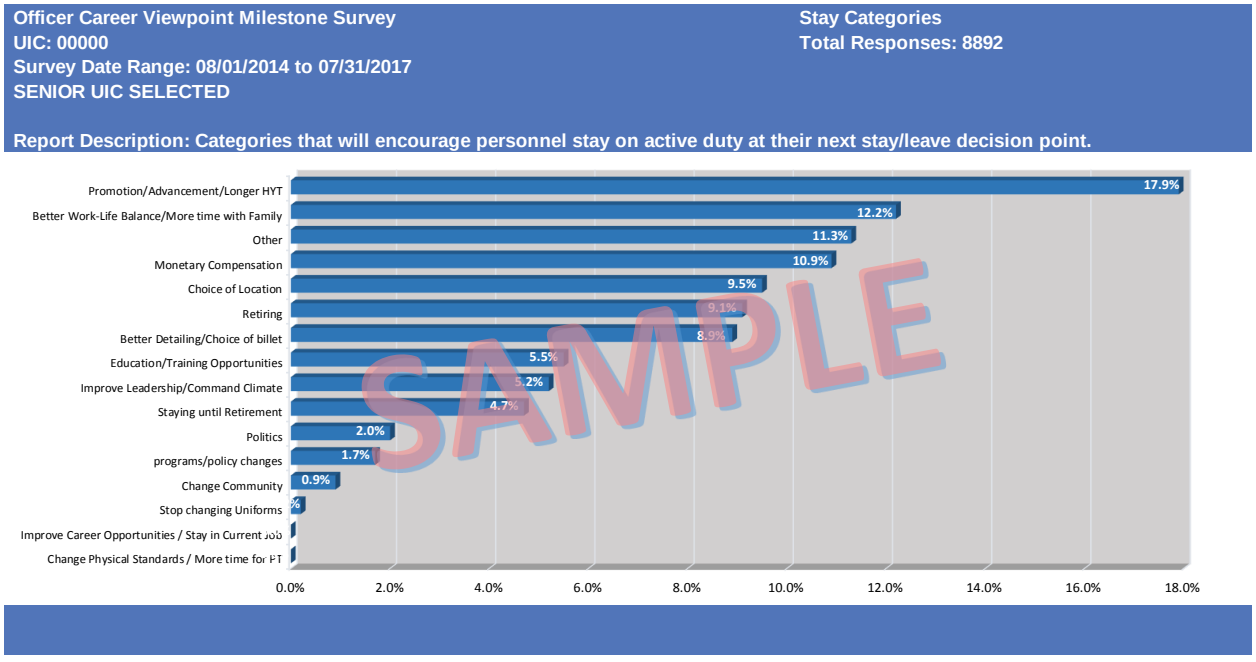


Figure 6. Sample DARTS Report Filtered for All Navy Officers

C. RETENTION SURVEY RESULTS

With five text comment questions from two active versions of the survey, the analysis focus in this section is on two questions that ask why personnel leave active duty and what will make them stay. The responses to these two questions are extracted from BO into an Excel file with the unique survey ID and several survey specific factors to allow the comments to be matched back to member demographics and military factors for further analysis.

The Milestone Survey contains the question "Using the space below, what can be done to encourage you to remain in the Navy on active duty when you are next required to make a stay/leave decision?" This comment box for this question has generated 13,781 responses. Using our methodology, we label and categorize these comments and include the results in a customized version of DARTS for use by OPNAV N1. Using DARTS, we find that Navy officers are leaving primarily because of civilian career opportunities and that increased promotion will help with retention.

In addition, OPNAV N1D has been instructed by CNO to provide research to assist in meeting the goal of eliminating gender bias in the Navy. One primary area of focus centers on the unrestricted line (URL) community that include of designators 11XX and 13XX. These designators correspond to the surface warfare, submarine, aviation, explosive ordinance disposal, and special warfare communities. These communities have an underrepresentation of female officers and retention of those that enter these communities is much lower than for their male counterparts. Utilizing DARTS and the Career Viewpoint Exit survey responses, we identify the top five reasons URL females are leaving the Navy as follows: 1. Civilian Career Opportunities, 2. Continued Service Not Authorized, 3. To Start or Focus on Family, 4. Retirement, 5. I do not fit in the Navy organization. These answers are informative, but they do not necessarily represent the target population of those early in their URL career, and they do not provide direct ways to increase retention.

Most URL officers have an MSR of between four and ten years, but the target population members are eligible to leave with about seven years of service. These

members take the Milestone Survey prior to six years of service, since it is distributed 13 months prior to their PRD or MSR.

To review the results for this population, we use DARTS to filter comments for these URL female officers. Our results show that over 30% of this group leaves active duty because they feel they spend too much time away from home while 25.6% indicate that they wish to start or focus more on their family. This same population comments that better work-life balance and more family time would encourage them to remain on active duty. Although bonuses are often pushed at this population, only 3% indicate that monetary compensation will encourage them stay. With these quantifiable results, retention policy makers are better able to review, modify, and create more relevant incentives to retain "our best sailors" while working within budget constraints and meeting fiscal year end strength and operational requirements.

An important consideration is to determine how the female URL results compare to their male counterparts as well as females who are not URL officers. This allows decision makers to determine if targeted changes are necessary for these communities or if there are larger scale issues that need to be addressed. Figure 7 shows a comparison of the primary comment topics that members indicate will encourage them to stay as filtered by the indicated demographics. The results show that there is a difference in view between female and male URL officers, with males looking for promotion and the ability to serve longer, while females want more personal time. The non-URL females have similar results, but the results confirm speculation that non-URL officers, (non-URL officers do not receive a bonus,) would like higher monetary compensation compared to URL officers who already receive a large bonus. The 9% difference in *better work-life balance / more time with family* is an indication that the URL community operational tempo is more of an issue and may need to be addressed to help retain female URL officers. DARTS, configured with the reasons leaving and encouragement to stay results, has been provided to OPNAV N1 for further assistance in responding to future tasking.

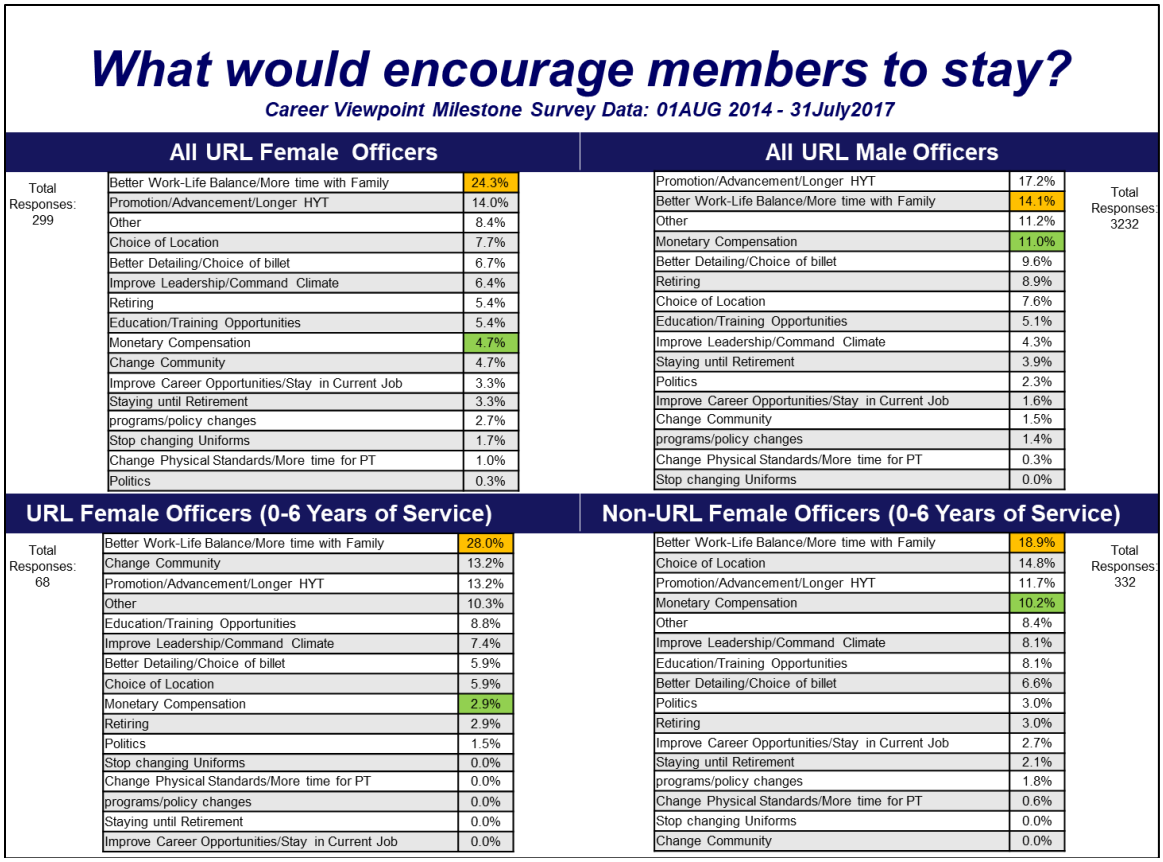


Figure 7. Retention Survey Result Comparison: Encouragement to Stay

IV. VALIDATION

Our methodology can provide two different levels of information. The summary level provides a quick analysis where primary topics are discovered quickly and can be used to make bins and keywords. The second level uses the topic summary to assign comments to topics for quantitative analysis. This chapter provides validation for both levels of our methodology.

A. TOPIC SUMMARY VALIDATION

For many surveys that have comments, the only purpose of the comment is to generate key topics from respondents and summarize their responses. The comment labels pull out prevalent topics and provide this summary with the aid of correlation plots and word clouds. This section confirms that our comment analysis methodology accomplishes this task by comparing an analysis of a survey using our methods to independent manual analysis of the same survey.

1. Survey Background

The Female Dress Uniform & Cover Survey was administered in March 2017 by OPNAV N1 to determine Fleet preference on type and design of female specific uniform items that were new or recently modified. The survey has eight topic based comment boxes that were read by OPNAV N1 staff over the course of two weeks to provide summaries for each comment. We compare these summaries to the primary topics using our method.

2. Labeling and Summary Comparison

Although all eight questions were reviewed and yield similar results, we discuss the validation for only one question, “What changes do you desire to the female SDB coat?” The correlation plot of salient label terms, displayed in Figure 8, includes the primary topics. The correlation plot of frequent label terms for this group of responses did not provide additional information and is not displayed.

Reviewing the correlation plot of salient terms, the thicker arc between two tokens indicates a stronger correlation between the tokens. It highlights several primary and relevant topics that respondents would like changed to the SDB coat including, "straight ribbons," "more form," "less bulky," "pockets," "material," "jumper style," [less like] "flight attendants," and "movement." Reviewing the word clouds with and without stop words (Figure 9), we find that the labels provide a more transparent summary of terms that could be displayed and confirm that the topics found with the correlation plot are reasonable. Our finding directly corresponds to the official 2017 summary provided by OPNAV N13XB that states

> Significant comments included alterations to the pockets. Navy females believe that pockets should be functional to allow for cell phones and should be straight to align with ribbons being parallel. Fewer comments noted that the SDB fit could be improved with an overall adjustment to the arm width and length to allow for more flexibility. The comfort of the material could be improved if it is more flexible and similar to the SDB slacks material.

This example illustrates that, with our method, summaries of text comments can be conducted in at least half the time it would take to read all the comments.

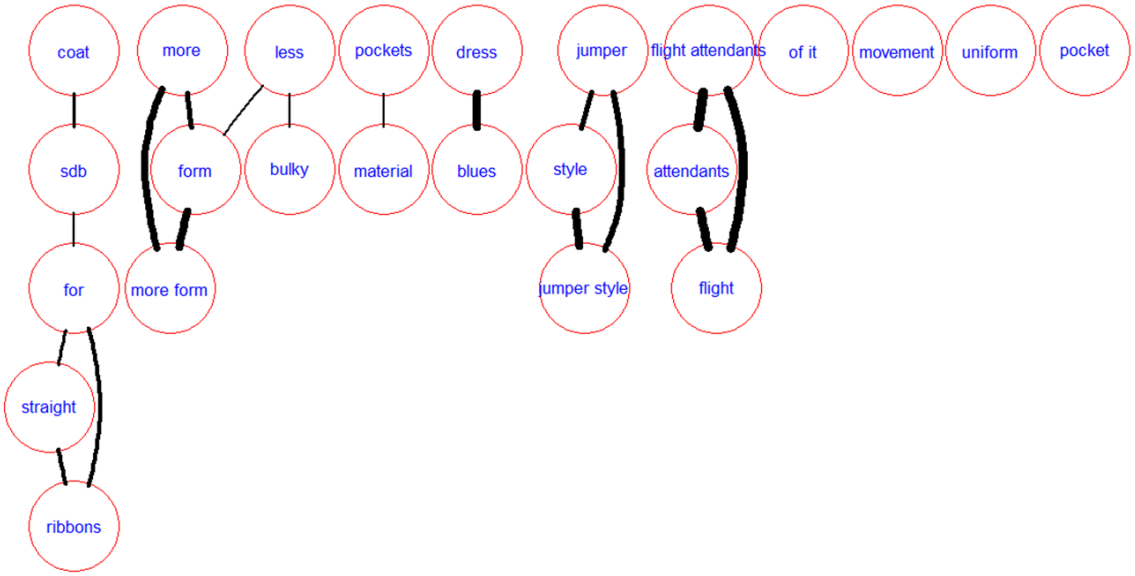


Figure 8. Correlation Plot of Salient Terms

Figure 9. Review of Word Clouds: Corpus without Stop Words, Labels with Stop Words, Labels without Stop Words

B. COMMENT BINNING VALIDATION

1. Expert Binning

We also validate our method using the Navy Retention Survey comments that ask why members leave active duty and what will encourage them to stay. These are manpower and retention based responses that require subject matter expertise related to the programs to make adequate binning recommendations. Five manpower experts were asked to participate in binning efforts with only two experts able to provide recommended bins due to the cumbersome nature of the task. The first is an OPNAV N1 analyst who works regularly with Navy retention and participated in the development of the Navy Retention Surveys. The other is a Navy lieutenant from the human resources community who has multiple tours as a human resources officer working in areas including recruiting, retention, and fleet support.

Each of the analysts is asked to individually read up to 200 comments. They were then asked to assign the comments to up to three of the provided bins created during the summary binning process and to rank their choices if they provided more than one. They were also asked to add an additional bin if they saw many similar comments that did not fit into provided bins.

2. Comparison

Of the 200 responses, 168 “reason leaving” comments and 44 “encouragement to stay” comments are binned by both experts. Only comments that are binned by both experts are used in the comparison to our comment analysis methodology results. Reviewing the “reason leaving” bins created by our comment analysis methodology, 30.4% of our binned comments are exact matches to both experts, 38.1% match the top bin of at least one expert, and 64.9% match one of the top three bins by either of the experts. It is important to note that our two experts only agreed on their first-choice bin for 46.2% of the comments.

The stay results indicate that 27.3% of our binned comments are exact matches to both experts, 45.5% match the top bin of at least one expert, and 56.8% match one of the top three responses by any of the experts with 43.2% of the primary ranked topic matching for the experts. The experts were provided nine bins for the encouragement to stay comments while an additional nine were identified through completion of our topic bin labeling process. Most of the comments labeled by our methodology in one of these additional nine bins were labeled as *Other* by our experts instead of adding a bin. To look for a better comparison, we remove *Other* responses that were binned by our methodology as one of the nine non-provided topics. There are 33 remaining comments of which, 36.4% of our binned comments are exact matches to both experts, 57.6% match the top bin of at least one expert, and 72.7% match one of the top three responses by any of the experts. This set had 51.5% of number 1 ranked bins matching for our experts.

These results indicate that there are different interpretations of every comment. Even two Navy manpower experts identified different bins for the same comment for more than half of the comments. Our results matching approximately 65% of at least one identified bin when bins are provided is comparable to levels attained by Chuang et al. (2012b) when labeling larger documents.

V. CONCLUSIONS AND FUTURE WORK

A. CONCLUSION

This methodology, as applied to short survey comments in a corpus independent manner, allows survey comments to be analyzed in a way not previously possible. This provides new objective results for the Navy where there has been limited quantitative evidence to justify retention bonuses or other retention policies. The use of DARTS with the Navy Retention Survey comments provides quantifiable reasons that targeted groups of members are leaving active service with indications of methods to better retain these sailors. Providing this tool to OPNAV N1 has answered questions from CNO concerning gender bias and reduces the number of individual communities that need to administer their own survey to justify a SRB. The filtering capabilities in DARTS proves invaluable for determining better ways to retain the Navy's members with critical skills and background.

Our method is generalizable beyond the realm of Navy retention. With over 65 surveys a year that contain short comments, survey administrators are always seeking tools to expedite the evaluation process and utilization of the comments. This research opens the door for a more effective feedback loop by analyzing more comments in a shorter period of time. With information from the comments, leadership can respond to sailor concerns and demonstrate the value of completing surveys. If sailors see that surveys make a direct impact, they will be more willing to complete them, continuing the chain of improving effective communication throughout the Navy.

Further our methods can be adopted to non-Navy surveys. The ability to adapt our approach by using a context specific reference corpus from any manual, document, or website allows the methodology to be applied to other surveys with a different set of jargon and acronyms. Because of this, our method is easily adaptable to any survey with topic based comments.

B. FUTURE WORK

This work provides a foundation for determining primary labels, topics, and bins from short-text comments using our comment analysis methodology. There are several avenues of future work that might expand our method to additional types of comments, and improve DARTS for a larger impact.

1. Allow Non-Consecutive Word Labels

Continued research is necessary to find a more robust algorithm that allows for labels to be applied from non-consecutive words on each comment. Human-generated labels often skip stop words, but the stop words are important to determining token variables used to score each token. Developing a method to utilize non-consecutive word *n*-grams will allow for increased matching to human responses and a greater understanding of important topics.

2. Opinion Based Comments

Opinion based comments are a type of comment that are frequently used in surveys. The use of these questions helps to avoid swaying a responder to one side or another of a discussion. These comments include *like* and *dislike* opinions of a topic and require an algorithm to determine a sentiment and a topic to correctly bin a comment. This is increasingly important for general "Please provide any additional comments" type questions that are not looking for answers to a specific question. Using sentiment in conjunction with identifying primary topics would improve analysis of these types of questions.

3. Automation of Initial Bin Key Creation

Automating the initial bin key creation would be an additional step to reduce subjectivity and provide a faster analysis with better replicability. With many similar topics that are not easily relatable outside of context, the bin creation and keyword list requires some analyst review, but creating a starting point with bins that could be grouped together or modified would provide another level of support. This could be

accomplished using lemmatization and synonym comparison techniques and the development of a comprehensive naval lexicon for use with Navy surveys.

4. Comprehensive Adaptation to DARTS

DARTS is a general tool that is adaptable to any single comment. This tool would be more valuable by including all comments from a specific survey and by building in a feature that allows the inclusion of quantitative questions as well. One tool for the complete analysis of a survey would be invaluable for the Navy and many other organizations that seek to quickly determine answers to important questions using quick turnaround surveys.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. CVSS DEMOGRAPHIC AND SERVICE ELEMENTS

UIC
UIC Command Name
Duty Status Code
Rating
Rate/Rank
Rate/Rank Description
Paygrade
Paygrade Description
Branch Class Code
Branch Class Abbreviation
Separation Program Designation Code
Separation Program Designation Short Description
Separation Program Designation Long Description
Officer Enlisted Indicator
Report Date
Accounting Category Code
Accounting Category Description
Active Duty Service Date
Active Commission Base Date
Sea Duty Commencement Date
Shore Duty Commencement Date
Special Program Indicator
Special Program Indicator Description
Program Enlisted Code
Program Enlisted Description
Contract Status Code
Current Enlistment Date
EOS Date
Soft EOS Date
EAOS Date
Extension Agreement No School
Extension Agreement School
SEAOS Date
Inoperable Extension Date
Sea Shore Code
Sea Shore Description
Naval Reserve Activity Code
Naval Reserve Activity Description
Manning Control Authority Code
Geographic Location Code
Global Location
Global Location Code
Geographic Broad Area
Geographic Broad Area Code
Geolocation City
Geolocation City Code
Geolocation County
Geolocation County Code
Geolocation Country
Geolocation Country Code
Geolocation State
Geolocation State Abbreviation
Geolocation State Code
Geolocation Region
Geolocation Coordinator
Geolocation Combination Code
Budget Submitting Office Code
Enlisted Management Community Code
Enlisted Management Community Code Description
Officer Designator Code
Officer Designator Desc
Primary NEC Code
Primary NEC Description
Secondary NEC Code
Secondary NEC Description
Previous Enlisted Indicator
Pay Entry Base Date
Projected Rotation Date
Date of Rank
Enlisted Warfare Designator Code
Enlisted Warfare Designator Description
Estimated Date of Loss to Navy
EDLN Reason Code
EDLN Reason Description
Commission Date
Military Service Requirement Date
Precedence Year Group
Year Group
Zone
Non Judicial Punishment Date
Individual Mobilization Status Code

Most Recent Exam Cycle
Most Recent Exam Date
Most Recent Exam Status
Detachment Estimated Date
Arrival Estimated Date
C-WAY Status Date
C-WAY NES Code
C-WAY Status
Pay Status Code
Marital Status
State(Home of Record)
State(Home of Record) Description
Dependent Status Code
Dependent Status Description
Number of Dependents
Family Co-location Indicator
Date of Birth
Age
Gender Code
Gender Description
Years of Education
Education Cert Code
Education Cert Description
Race Code
Race Description
Ethnic Code
Ethnic Description

APPENDIX B. SAMPLE RE-CREATION OF A PEOPLESOFT UIC LEVEL CAREER VIEWPOINT MILESTONE SURVEY REPORT

The following is a user created, Excel report that is shown to illustrate the standard report that is available in the CVSS CIMS access in NSIPS. The PeopleSoft, pre-built report is protected by the Privacy Act of 1974. This "UIC level" re-creation contains the same formatted representation as the original report, but is available for public release.

CAREER VIEWPOINT MILESTONE SURVEY SUMMARY REPORT
BUREAU OF NAVAL PERSONNEL
2014-08-01 to 2017-08-01

Core Question Summary

Overall Response Rate: 51.67%

Enlisted

Response Rate: 51.43%

Career Intentions

50.0%	To remain in the Navy on active duty until I am eligible to retire (or longe
25.0%	I'm not sure
16.7%	I am eligible for retirement but I intend to stay on active duty
8.3%	Leave active duty as soon as I can

Top 5 Reasons to Stay

91.7%	Monetary compensation & retirement
83.3%	Medical/Dental benefits (member and/or family)
75.0%	Other benefits (leave, education, commissary, NEX)
58.3%	Promotion/Advancement opportunities
58.3%	Work-life balance (operational demand, sea duty)

Top 5 Reasons to Leave

58.3%	Impact on family (support, moving, child care)
50.0%	Career assignments (options, member control)
50.0%	Command climate (previous and current commands)
41.7%	Leadership (All Navy and command)
25.0%	Work-life balance (operational demand, sea duty)

Officer

Response Rate: 51.85%

Career Intentions

57.9%	To remain in the Navy on active duty until I am eligible to retire (or longe
31.6%	I am eligible for retirement but I intend to stay on active duty
10.5%	I'm not sure

Top 5 Reasons to Stay

94.7%	Medical/Dental benefits (member and/or family)
78.9%	Monetary compensation & retirement
73.7%	Promotion/Advancement opportunities
73.7%	Other benefits (leave, education, commissary, NEX)
68.4%	Current job satisfaction

Top 5 Reasons to Leave

52.6%	Impact on family (support, moving, child care)
47.4%	Career assignments (options, member control)
42.1%	Work-life balance (operational demand, sea duty)
31.6%	Civilian job opportunities
21.1%	Command climate (previous and current commands)

Detailed Question Summary

Overall Response Rate: 50.00%

Enlisted

Response Rate: 60.00%

Top 10 Reasons to Stay

75.0%	Overall pay
75.0%	Retirement pay and benefits
75.0%	Service member's medical benefits
66.7%	Base pay
66.7%	Tuition Assistance (TA) benefits
66.7%	Overall value of benefits
58.3%	Competence of co-workers
58.3%	Competence of supervisors
58.3%	Basic Allowance for Housing (BAH)
58.3%	Quality of medical care for member's family

Top 10 Reasons to Leave

58.3%	Separation from family and friends
41.7%	The balance between work and personal time
41.7%	The red tape required to complete tasks
41.7%	The impact of being in the Navy on family
41.7%	The impact of PCS moves on a spouse's career
41.7%	The impact of PCS moves on children
33.3%	Variety of job choices available
33.3%	Schedule changes and unpredictability
33.3%	Impact of geographic location on spouse's career
33.3%	Work time available to keep physically fit

Officer

Response Rate: 45.00%

Top 10 Reasons to Stay

83.3%	Retirement pay and benefits
72.2%	Overall pay
72.2%	Service member's dental benefits
66.7%	Meaningfulness of the work service member does
66.7%	Base pay
66.7%	Basic Allowance for Housing (BAH)
66.7%	Service member's medical benefits
61.1%	Availability of medical care for member's family
61.1%	Location of medical care for member's family
61.1%	Quality of medical care for member's family

Top 10 Reasons to Leave

44.4%	The red tape required to complete tasks
44.4%	The impact of being in the Navy on family
44.4%	Children's education
38.9%	Variety of job choices available
38.9%	Schedule changes and unpredictability
38.9%	The balance between work and personal time
38.9%	Separation from family and friends
38.9%	The impact of PCS moves on children
33.3%	Control of orders to desired geographical location
33.3%	Time spent deployed

Policy Question Summary

Enlisted

Navy Policy Most Influential Reason to Stay

8.3% Reenlistment Opportunities

Navy Policy Most Influential Reason to Leave

8.3% Advancement Opportunity

Officer

Navy Policy Most Influential Reason to Stay

5.6% Promotion Opportunity

Navy Policy Most Influential Reason to Leave

11.1% Promotion Opportunity

APPENDIX C. SAMPLE RE-CREATION OF A BUSINESSOBJECTS ALL NAVY OFFICER CAREER VIEWPOINT EXIT SURVEY REPORT

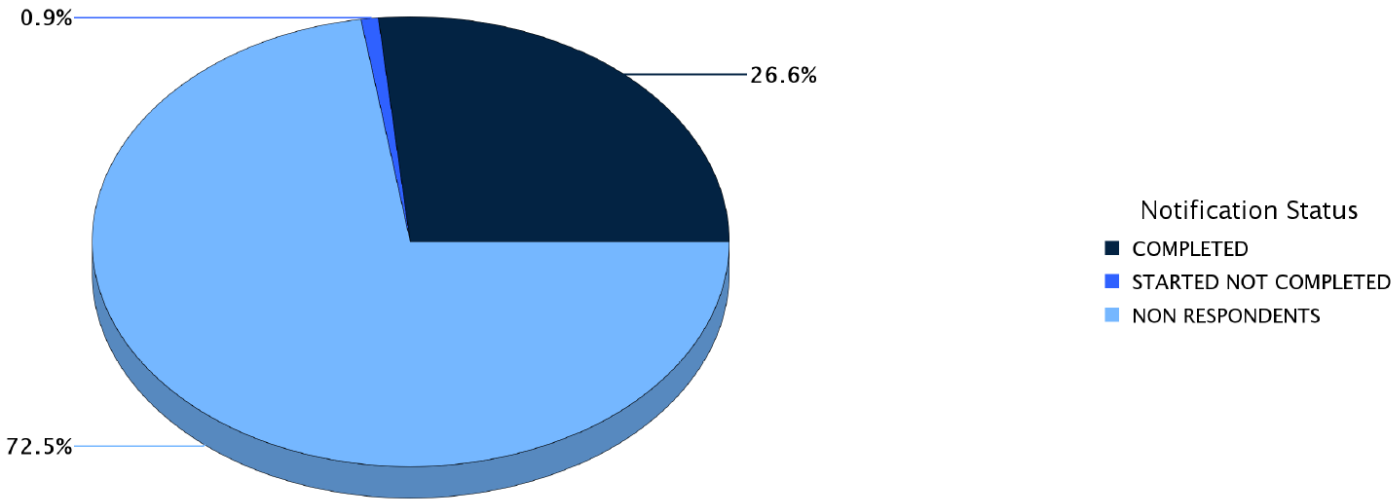
The following is an ad hoc, user-created, CVSS BusinessObjects report that is shown to illustrate the standard report that is available in the CVSS BuisinessObjects access in NSIPS. The pre-built report is classified as "For Official Use Only" since it can be filtered for respondent demographics. This "All Navy" re-creation contains the same graphical representation as the original report but is available for public release.

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

PARTICIPATION REPORT
Total Survey Notifications 4,079
All Personnel

Survey participation status of service members who were requested to take the survey.



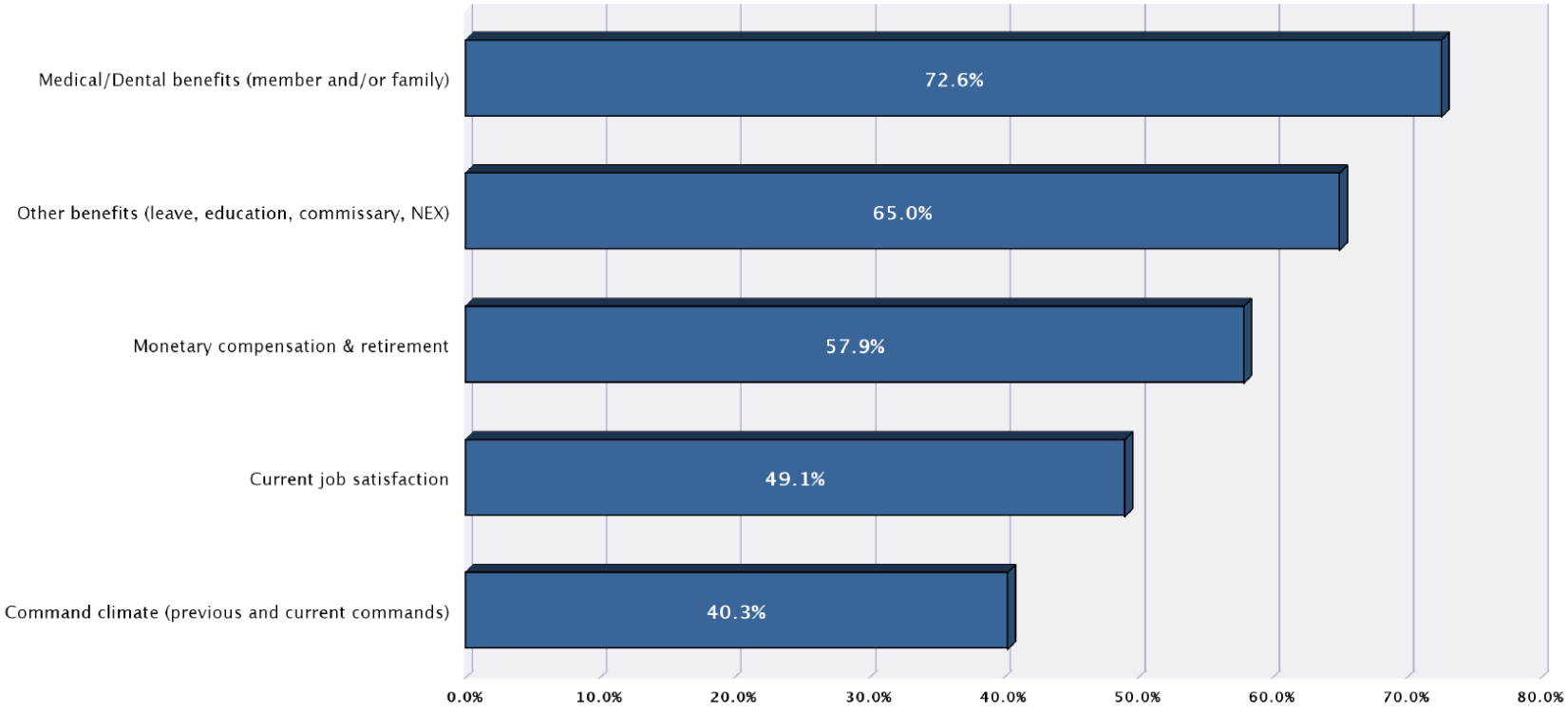
SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

TOP 5 CORE INFLUENCERS TO STAY
Total Survey Respondents 1,121
All Personnel

Top 5 influencers to stay in the Navy based on the required core questions.



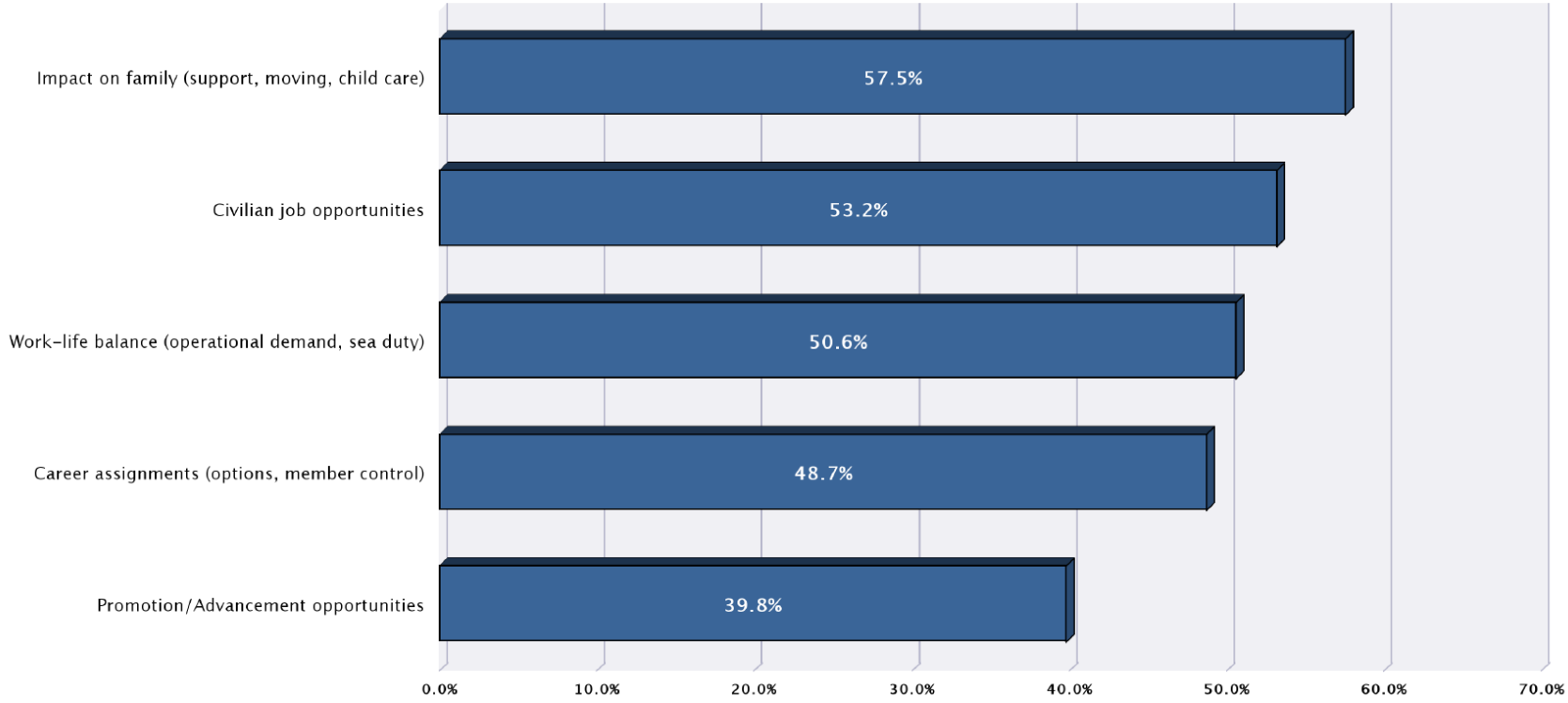
SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

TOP 5 CORE INFLUENCERS TO LEAVE
Total Survey Respondents 1,121
All Personnel

Top 5 influencers to leave the Navy based on the required core questions.



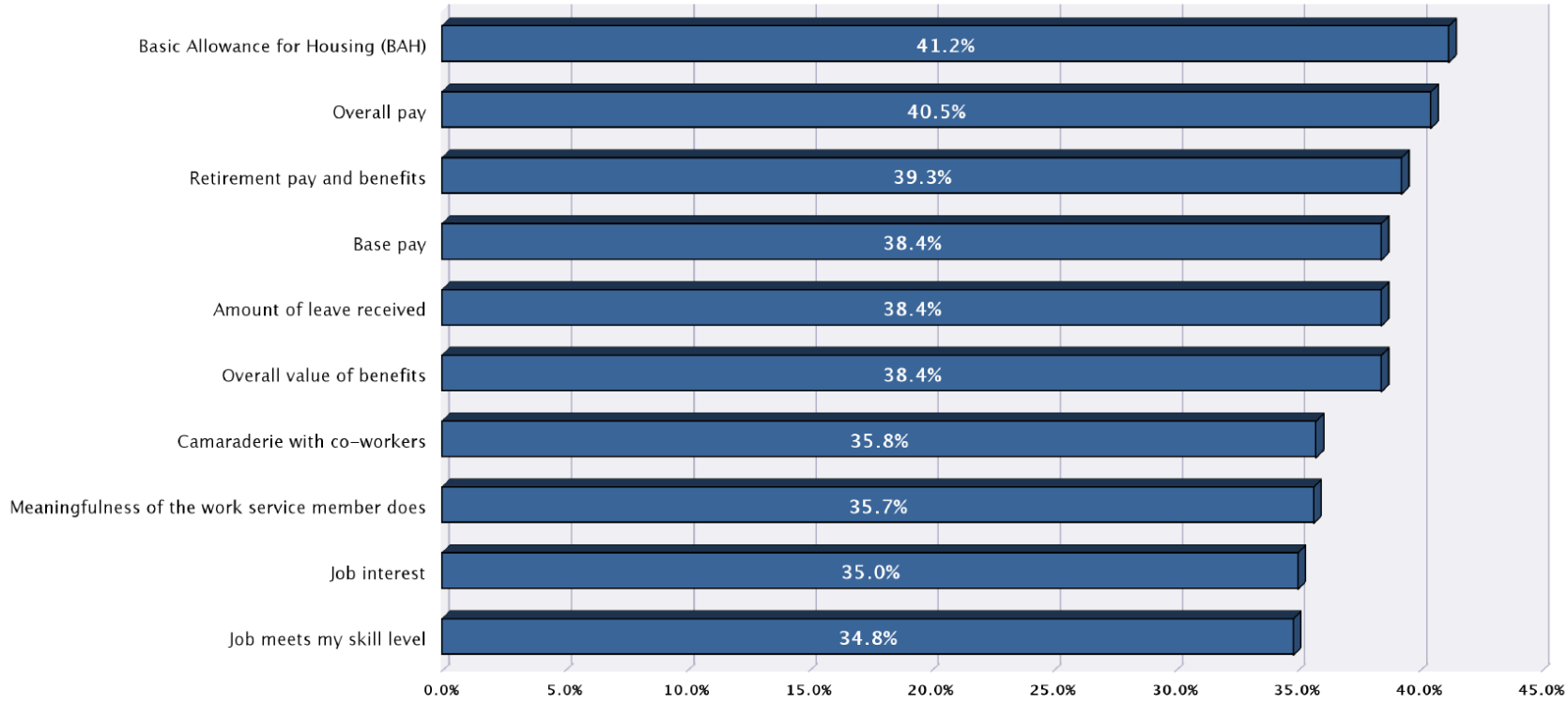
SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

TOP 10 INFLUENCERS TO STAY
Total Survey Respondents 1,085
All Personnel

Top 10 influencers to stay in the Navy based on the non-required questions.



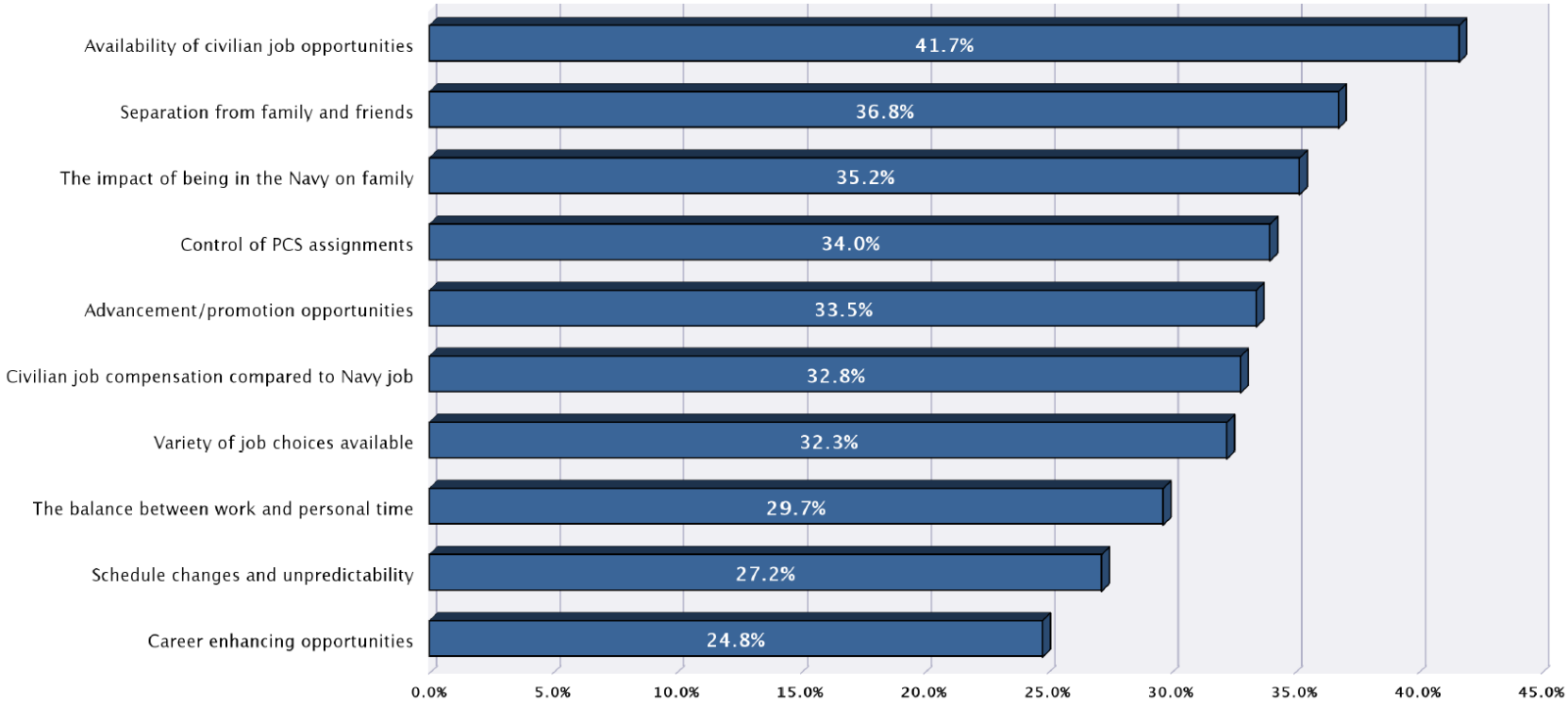
SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

TOP 10 INFLUENCERS TO LEAVE
Total Survey Respondents 1,085
All Personnel

Top 10 influencers to leave the Navy based on the non-required questions.



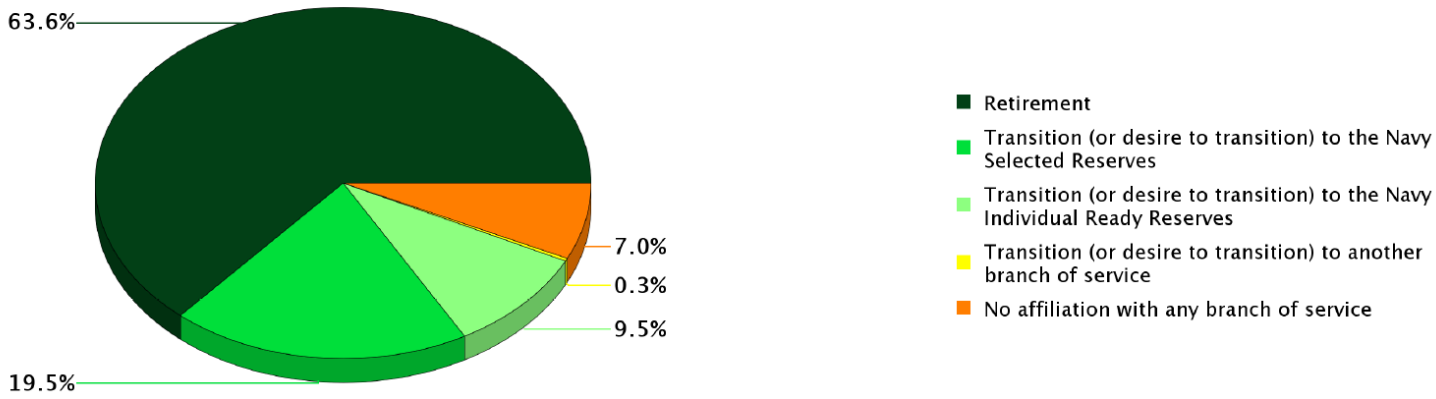
SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

NAVY CAREER INTENTIONS
Total Survey Respondents 1,121
All Personnel

Career intentions for continued service in the Navy based on respondents frequency to a required question.



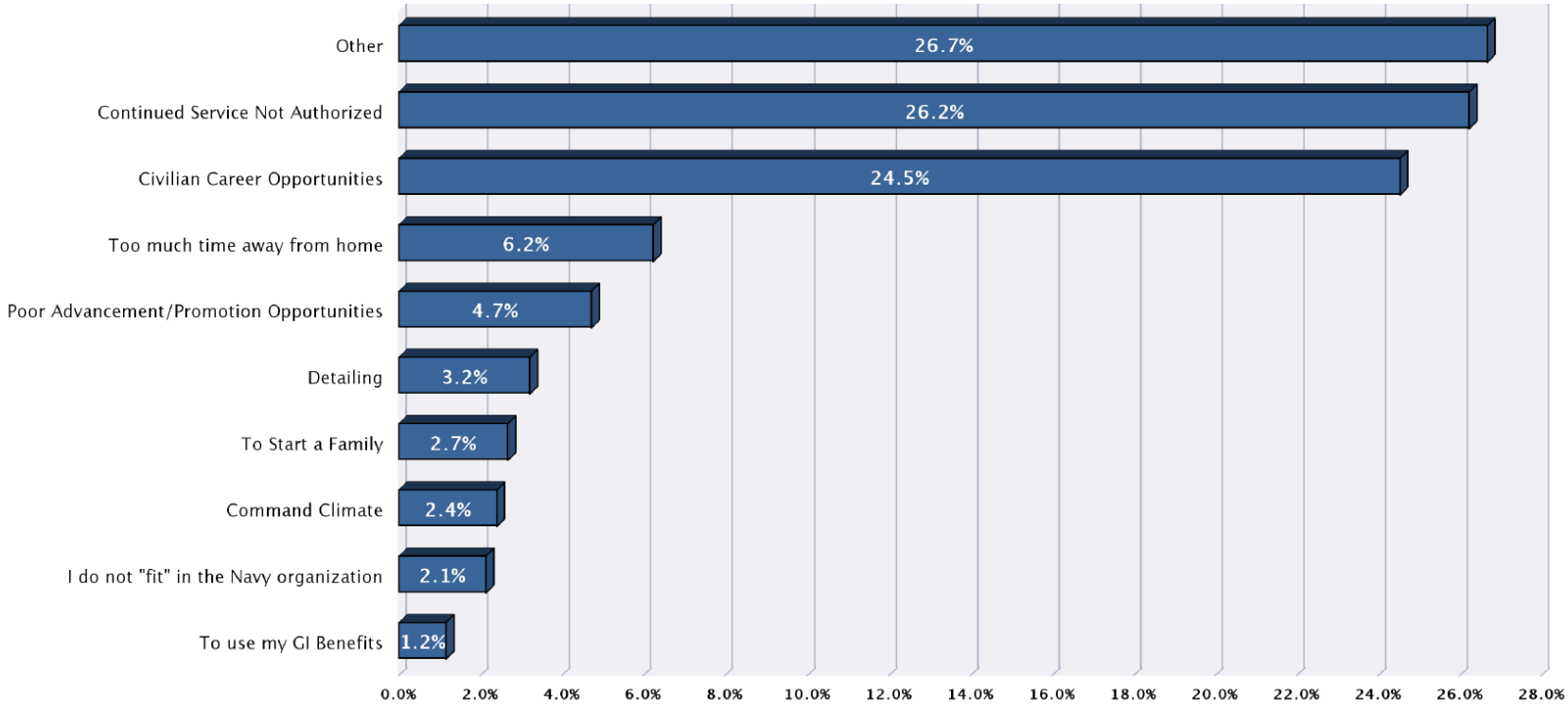
SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

Run Date: 09/28/2017 05:07 PM
Current User: N1291225086S0020

Officer Career Viewpoint Exit Survey
Survey ID: 10002
UIC: 00000
Survey Date Range: 9/1/15 to 8/31/16
SENIOR AND SUBORDINATE UIC(s) SELECTED

PRIMARY REASON FOR LEAVING THE NAVY
Total Survey Respondents 1,121
All Personnel

Primary reason for leaving active duty in the Navy based on respondents frequency to a required question.



SENIOR AND SUBORDINATE UIC(s) SELECTED for UIC(s):
00000

APPENDIX D. PREPROCESSING SUBSTITUTIONS AND CONTRACTIONS

Find	Replace
higher tenure	hyt
high year tenure	hyt
high tenure	hyt
high tenor	hyt
higher year tenure	hyt
higher tenor	hyt
failed officer selection	fos
failure to select	fos
failed to select	fos
fail to select	fos
prt	pfa
pfa	pfa
bca	pfa
parent	family
children	family
child	family
son	family
daughter	family

Find	Replace
wife	family
husband	family
spouse	family
mother	family
mom	family
dad	family
father	family
kid	family
kids	family
job	career
employment	career
&	and
sailor	.
school	education
college	education
schooling	education
gi bill	education
degree	education

Contraction	Replace
‘cause	because
‘tis	it is
‘twas	it was
ain’t	am not
aren’t	are not
can’t	can not
could’ve	could have
couldn’t	could not
didn’t	did not
doesn’t	does not
don’t	do not
hasn’t	has not
he’d	he would
he’ll	he will
he’s	he is
how’d	how did
how’ll	how will
how’s	how is
I’d	I would
I’ll	I will
I’m	I am
I’ve	I have
isn’t	is not
it’ll	it will
it’s	it is
let’s	let us
might’ve	might have
mightn’t	might not
must’ve	must have
mustn’t	must not
shan’t	shall not
she’d	she would
she’ll	she will
she’s	she is
should’ve	should have

Contraction	Replace
shouldn’t	should not
that’ll	that will
that’s	that is
there’ll	there will
there’s	there is
they’d	they would
they’ll	they will
they’re	they are
they’ve	they have
wasn’t	was not
we’d	we would
we’ll	we will
we’re	we are
we’ve	we have
weren’t	were not
what’s	what is
what’d	what did
when’d	when did
when’ll	when will
when’s	when is
where’ll	where will
where’s	where is
who’d	who would
who’ll	who will
who’s	who is
why’d	why did
why’ll	why will
why’s	why is
won’t	will not
would’ve	would have
wouldn’t	would not
you’d	you would
you’ll	you will
you’re	you are
you’ve	you have

APPENDIX E. ENCOURAGEMENT TO STAY TOPIC BIN KEY USING REGULAR EXPRESSIONS

not applicable	77	twenty years	2
NA’	77	careerist	2
ntr	77	rat.{0,7}chang	9
not.{0,12}sure	77	chang.{0,7}rate	9
not know	77	different.{0,}rat	9
no idea	77	switch	9
no.{0,12}comment	77	cross rat	9
staying	77	crossrat	9
going.{0,12}stay	15	lateral	9
intend.{0,12}stay	15	change job	9
plan.{0,12}stay	15	change designator	9
like.{0,12}stay	15	change communit	9
want.{0,12}stay	15	conver	9
remain.{0,12}active	15	transfer out	9
over 2	15	another communit	9
over 3	15	desig.{0,12}communit	9
3.{0,10}year	15	desig.{0,12}warfar	9
2.{0,10}year	15	leader	8
will.{0,10}remain	15	command climate	8
nothing	14	communication	8
none	14	coc	8
not for me	14	chain of command	8
policy	13	education	7
program	13	school	7
chang.{0,12}polic	13	masters	7
current rate	12	tuition	7
stay rate	12	learn	7
commission	11	training	7
sta 21	11	ta benefit	7
sta21	11	assignment	6
cwo	11	detail	6
ldo	11	billet	6
officer	11	job	6
retir	2	orders	6
am.{0,7}stay	2	famil	5
staying.{0,7}in	2	shore	5
reenlist	2	balance	5
renlist	2	short	5
decid.{0,7}stay	2	deploy	5

sea	5	pick.{0,9}up.{0,7}e	3
work hours	5	mak.{0,7}e	3
monetary	4	location	1
compensation	4	duty station	1
money	4	closer	1
bonus	4	pcs	1
srb	4	station	1
pay	4	next duty	1
paid	4	geographic	1
salary	4	port	1
wage	4	not.{0,7}transfer	1
bah	4	homestead	1
advanc	3	region	1
promot	3	physical	16
eval	3	pfa	16
hyt	3	prt	16
fitrep	3	weight	16
chief	3	bca	16
cpo	3	change.{0,12}uniform	17
next PG	3	stop.{0,16}uniform	17
mak.{0,7}rank	3	politic	18
pick.{0,9}up.{0,7}rank	3	bureaucracy	18
select.{0,12}e	3	admin	18
select.{0,7}rank	3	career.{0,12}opp	12

LIST OF REFERENCES

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Chen, P., Zhang, N. L., Liu, T., Poon, L. K. M., Chen, Z., & Khawar, F. (2017). Latent tree models for hierarchical topic detection. *Artificial Intelligence*, 250, 105–124. doi:10.1016/j.artint.2017.06.004

Chuang, J. (2013). *Designing visual text analysis methods to support sensemaking and modeling* (Doctoral dissertation). Retrieved from https://nlp.stanford.edu/~manning/dissertations/Chuang-Jason-dissertation.pdf

Chuang, J., Manning, C., & Heer, J. (2012a). Termite: Visualization techniques for assessing textual topic models. ACM. doi:10.1145/2254556.2254572

Chuang, J., Manning, C., & Heer, J. (2012b). Without the clutter of unimportant words. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3), 1–29. doi:10.1145/2362364.2362367

Department of the Navy. (2002). *Naval military personnel manual* (MILPERSMAN) (NAVPERS 15560D). Millington, TN: Bureau of Naval Personnel.

Department of the Navy. (2011). United States Navy uniform regulations (NAVPERS 15665I). Millington, TN: Bureau of Naval Personnel.

Director, Military Personnel Plans and Policy Division (N13). (2013, June 26). Career viewpoint survey items [Memorandum]. Washington, DC: Office of the Chief of Naval Operations.

Feinerer, I. & Hornik, K. (2017). tm: Text Mining Package. Retrieved from https://CRAN.R-project.org/package=tm

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. doi:10.18637/jss.v040.i13

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, *24*(2), 225–232. doi:10.1007/s00180-008-0119-7

Hornik, K. (2016). OpenNLP: Apache OpenNLP tools interface. Retrieved from https://CRAN.R-project.org/package=openNLP

Justeson, J., & Katz, S. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. doi:10.1017/S1351324900000048

Kim, S., Medelyan, O., Kan, M., & Baldwin, T. (2010). SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation*, Sweden, 21–26, Retrieved from http://www.aclweb.org/anthology/S10-1004/

Kullback, S., & Leibler, R. (1951). On the information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. Retrieved from http://www.jstor.org/stable/2236703

Lockheed Martin, NSIPS Program. (2013). Software design description for the navy standard integrated personnel system, Career Viewpoint Surveys and Studies (SG26OY1010B). New Orleans, LA: Program Executive Office, Enterprise Information System (PEO-EIS).

MAX.gov. Surveys. (2017). *2017 Female dress uniform & cover survey*. Retrieved from https://survey.max.gov

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 490–499). ACM.

Navy Standard Integrated Personnel System, CIMS. (n.d.). *Career viewpoint retention survey*. Retrieved from https://nsipsprod.nmci.navy.mil/

Navy Standard Integrated Personnel System, CVSS. (n.d.). *Career viewpoint retention survey*. Retrieved from https://nsipsprod.nmci.navy.mil/

Nitin, G. I., Swapna, G., & Shankararaman, V. (2015). Analyzing educational comments for topics and sentiments: A text analytics approach. *IEEE*. doi:10.1109/FIE.2015.7344296

Ooms, J. (2017). pdftools: Text Extraction and Rendering of PDF Documents. Retrieved from R package 1.2. https://CRAN.R-project.org/package=pdftools

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from https://www.R-project.org/

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach.* Sebastopol, CA: O'Reilly.

United States Navy. (n.d.). Chief of Naval Personnel. Retrieved October 31, 2016, from http://www.navy.mil/navydata/leadership/cnp_resp.asp

Zhai, C., & Massung, S. (2016). *Text data management and analysis: A practical introduction to information retrieval and text mining.* New York, NY: ACM Books.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California